

FUNDAMENTALS OF OPTICS

,

FUNDAMENTALS OF OPTICS

BY
FRANCIS A. JENKINS

*Professor of Physics
University of California*

AND
HARVEY E. WHITE

*Professor of Physics
University of California*

SECOND EDITION
Third Impression

McGRAW-HILL PUBLISHING COMPANY LTD.
NEW YORK LONDON TORONTO

1923

- To
Raymond Thayer Birge

PREFACE TO THE SECOND EDITION

As compared to its predecessor, "Fundamentals of Physical Optics," the present text embodies extensions in two directions. The first ten chapters on geometrical optics have been added at the suggestion of a number of users who desire to include at least a brief review of the principles of this subject. At the University of California geometrical and physical optics are now presented in separate courses, and therefore the former subject is here carried somewhat beyond the elementary level. Although we begin with first principles, these are introduced in a more general form than is customarily used for beginners. The second extension is the inclusion of a final chapter on the quantum behavior of light. The unity of subject matter which was achieved in "Fundamentals of Physical Optics" by covering exclusively wave optics has thereby been sacrificed, but experience leads us to believe that this drawback will be more than compensated by an increase in the general usefulness of the book.

The field of optics has continued to show a vigorous growth in recent years. New and important applications of old principles have appeared, and doubtless others are yet to come. Developments such as the phase-contrast microscope, directional radar, nonreflecting films, the electron microscope, and the Schmidt camera testify to the practical importance of optics as a field of study. Although many of the new discoveries are either too specialized or too complex to warrant a detailed treatment in a textbook devoted to fundamentals, they have been included wherever they furnish apt illustrations of the principles discussed. The former text on physical optics has been carefully scrutinized, not only from the standpoint of inserting such examples, but also from that of improving the clarity and rigor of its presentation. In this connection the authors must acknowledge their deep obligation to a number of persons who have taken great pains to communicate their suggestions and criticisms. Besides Professor R. T. Birge and other members of the Department of Physics at Berkeley, those who have been most helpful in this connection are Professors J. W. Eils, L. B. Heilprin, P. Kirkpatrick, C. F. Meyer, W. M. Preston, E. M. Purcell, G. D. Rochester, W. W. Sleator, and M. W. Zemansky.

The device of including brief descriptions of experimental illustrations, set off the main text by horizontal rulings, is not as useful nor as appro-

priate in geometrical as in physical optics. Instead, we have used in the early chapters the same method of segregation for the solutions of numerical examples.

The lists of problems at the ends of chapters have been completely changed and many new problems added. Answers for the even-numbered ones are given at the back of the book.

As regards laboratory experiments, we have not thought it advisable to include descriptions of these, because of lack of space and of the fact that the available equipment often varies greatly in different places. The laboratory manuals now in use at the University of California may be obtained by writing the Associated Students' Store, Berkeley, requesting "Laboratory Exercises in Geometrical Optics" and "Laboratory Exercises in Physical Optics," both by R. S. Minor and H. E. White.

FRANCIS A. JENKINS
HARVEY E. WHITE

PREFACE TO THE FIRST EDITION

This textbook is intended for use in an advanced undergraduate course in optics. It is assumed that the student has completed a thorough course in elementary physics and is familiar with the methods of the calculus. We have presented the material in such a way, however, that the book may be used in classes in which some students do not have the above mathematical preparation. Thus, wherever possible, the mathematical derivations are supplemented by simple graphical or vector treatments of the problem. The applications of calculus have been purposely made very brief, but are always included for the benefit of those students with a mathematical turn of mind. The main emphasis is placed on the physical explanation of the various phenomena, which we believe is most successfully accomplished in the present subject by the use of graphical methods. For this reason a large number of illustrations have been prepared with considerable care to have them as exact as possible.

We have deliberately restricted the subject matter to rather narrow limits. Thus, on the one hand, we have included no geometrical optics. The knowledge of this subject gained in an elementary physics course is ample for an understanding of the material of this book. On the other hand, no systematic discussion of the quantum theory and its applications to spectra and atomic structure has been given, even though this is an essential part of the subject of physical optics as the term is generally understood. It would have been impossible to include an adequate account of this field without a considerable increase in the size of the book. We have therefore limited ourselves strictly to the so-called classical physical optics, or wave optics. This has been necessary in order that there should be room for a sufficiently detailed consideration of our subject.

But there is a more fundamental reason for this limitation than the mere exigencies of space. The complementary character of the wave and quantum aspects of light, which is an essential part of the modern theory, reveals these as two equally important, but quite distinct, fields of study. In covering only the one field, the book achieves a unity which would be lost by the inclusion of a necessarily brief account of the other field. The usual procedure in an introductory presentation of light has been to develop the wave theory first, and afterward to describe some of the quantum phenomena requiring the particle theory. The dilemma in which we are left concerning the true nature of light is then emphasized in such a way as to leave the impression that ultimately one or the other of these theories will prove to be correct. It seems to us that the time

has come to adopt the point of view emphasized by the quantum mechanics, namely, that the wave and particle properties of light are merely two different aspects of the same thing, and that one will probably never be more important than the other. These two aspects are to be regarded as complementary rather than as antagonistic. Although the acceptance of this point of view requires a fundamental change in our ideas as to what constitutes an "explanation" of a phenomenon, the thoughtful student should certainly be given the benefit of this newer outlook.

The dual character of the present theory of light and matter leads to a logical way of dividing the subject matter into two parts. On the one hand classical mechanics, the mechanics of particles, corresponds to the quantum picture of light and to geometrical optics. On the other hand the wave mechanics corresponds to wave optics. In confining ourselves to the latter field, we are covering the subject of "physical optics" in the sense of classical physics only. In our opinion the quantum aspects of light, which are apparently so sharply divided from the wave aspects, are best presented in a separate course. If it is desired to include them in the same course, reference should be made to other books in which a fairly complete treatment of the quantum theory is given. To be sure, it was not necessary or desirable to omit all mention of the quantum aspects in the present book. In the later chapters, which deal with the interaction between light and matter, we have been careful to point out the shortcomings of the wave picture, and the necessity of turning to the quantum theory for a complete explanation.

The most beautiful and striking experiments in physics are to be found in the field of physical optics. Hence it is very desirable that as many as possible of these be shown to the class or performed by the students themselves. Descriptions of many demonstration experiments are given throughout the text; these are set off from the text itself by horizontal lines. The laboratory work accompanying the course now given at the University of California is described in "Laboratory Experiments in Physical Optics," by R. S. Minor and H. E. White.

In writing this text we have had free access to the lecture notes used by Professor R. T. Birge in his advanced course on physical optics, and from these we have taken some of the explanations and drawings used in the more involved phases of the subject. We are also deeply indebted to Professor Birge for reading the entire manuscript and for making numerous valuable suggestions in regard to it. We wish to express our sincere thanks to Professor R. S. Minor for the ruling of the various special diffraction gratings used by us in obtaining the photographs in Figs. 6A, 6E, 7A, and 7F.

FRANCIS A. JENKINS
HARVEY E. WHITE

CONTENTS

	PAGE
PREFACE TO THE SECOND EDITION.	vii
PREFACE TO THE FIRST EDITION.	ix
PART I. GEOMETRICAL OPTICS	
1. Light Rays.	3
2. Plane Surfaces	17
3. Thin Lenses.	36
4. Spherical Surfaces.	51
5. Thick Lenses	65
6. Spherical Mirrors	80
7. The Effects of Stops.	94
8. Ray Tracing	113
9. Lens Aberrations	123
10. Optical Instruments	155
PART II. PHYSICAL OPTICS	
11. Light Waves	179
12. The Superposition of Waves	203
13. Interference of Two Beams of Light	225
14. Interference Involving Multiple Reflections	254
15. Fraunhofer Diffraction by a Single Opening	279
16. The Double Slit.	303
17. The Diffraction Grating	320
18. Fresnel Diffraction	347
19. The Velocity of Light	378
20. The Electromagnetic Character of Light	406
21. Sources of Light and Their Spectra	420
22. Absorption and Scattering.	444
23. Dispersion	462
24. The Polarization of Light.	486
25. Double Refraction.	507
26. Interference of Polarized Light	527
27. Optical Activity.	543
28. Reflection	560
29. Magneto-optics and Electro-optics	589
PART III. QUANTUM OPTICS	
30. Photons	611
ANSWERS TO EVEN-NUMBERED PROBLEMS	627
INDEX.	631

Part I
GEOMETRICAL OPTICS

CHAPTER 1

LIGHT RAYS

Optics, the study of light, is conveniently divided into three fields, each of which requires a markedly different method of theoretical treatment. These are (a) *geometrical optics*, which is treated by the method of light rays, (b) *physical optics*, which is concerned with the nature of light and involves primarily the theory of waves, and (c) *quantum optics*, which deals with the interaction of light with the atomic entities of matter and which for an exact treatment requires the methods of quantum mechanics. This book deals almost entirely with (a) and (b), although some of the salient features of (c) will be outlined in the last chapter. These fields might preferably be called macroscopic, microscopic, and atomic optics as giving a more specific indication of their domains of applicability. When it is a question of the behavior of light on a large scale, the representation by means of rays is almost always sufficient.

1.1. Concept of a Ray of Light. The distinction between geometrical and physical optics appears at once when we attempt by means of

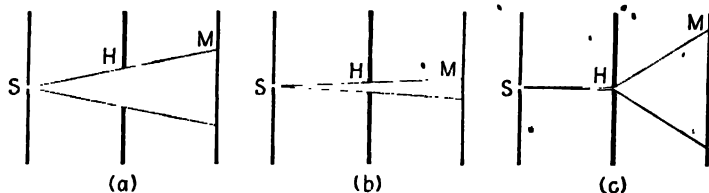


FIG. 1A. Attempt to isolate a single ray of light.

diaphragms to isolate a single ray of light. In Fig. 1A let S represent a source of light of the smallest possible size, a so-called *point source*. Such a source is commonly realized by focusing the light from the white-hot positive pole of a carbon arc on a metal screen pierced with a small hole. If another opaque screen H provided with a much larger hole is now interposed between S and a white observing screen M [Fig. 1A(a)], only the portion of the latter lying between the straight lines drawn from S will be appreciably illuminated. This observation forms the basis for saying that light is propagated in straight lines called *rays*, since it can

* The concentrated arc lamp to be described in Sec. 21.2 also furnishes a very convenient way of approximating a point source.

be explained by assuming that only the rays not intercepted by H reach the observing screen. If the hole in H is made smaller, as in (b) of the figure, the illuminated region shrinks correspondingly, so that one might hope to isolate a single ray by making it vanishingly small. Experiment shows, however, that at a certain width of H (a few tenths of a millimeter) the bright spot begins to widen again. The result of making the hole exceedingly small is to cause the illumination, although it is very feeble, to spread over a considerable region of the screen [Fig. 1A(c)].

The failure of this attempt to isolate a ray is due to the process called *diffraction*, which also accounts for a slight lack of sharpness of the edge of the shadow when the hole is wider. Diffraction is a consequence of the wave character of light and will be fully discussed in the section on physical optics. It becomes important only when small-scale phenomena are being considered, as in the use of a fine hole or in the examination of the edge of the shadow with a magnifier. In most optical instruments, however, we deal with fairly wide beams of light and the effects of diffraction can usually be neglected. The concept of light rays is then a very useful one because *the rays show the direction of flow of energy in the light beam*.

1.2. Laws of Reflection and Refraction. These two laws were discovered experimentally long before their significance was understood, and together they form the basis of the whole of geometrical optics. They may be derived from certain general principles to be discussed later, but

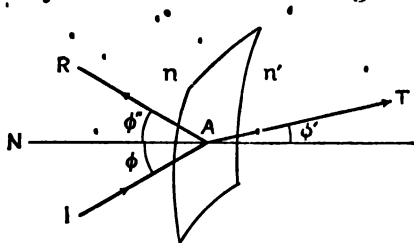


FIG. 1B. Reflection and refraction of a ray at a boundary.

for the present we shall merely state them as experimental facts. When a ray of light strikes any boundary between two transparent substances in which the velocity of light is appreciably different, it is in general divided into a reflected ray and a refracted ray. In Fig. 1B let IA represent the incident ray, and let it make the angle ϕ with NA , the normal or perpendicular to the surface at A . ϕ is called the *angle of incidence* and the plane defined by IA and NA is called the *plane of incidence*.

The law of reflection may now be stated as follows:

The reflected ray lies in the plane of incidence, and the angle of reflection equals the angle of incidence.

That is, IA , NA , and AR are all in the same plane and

$$\phi'' = \phi \quad (1a)$$

The *law of refraction*, usually called Snell's law after its discoverer,* states that

The refracted ray lies in the plane of incidence, and the sine of the angle of refraction bears a constant ratio to the sine of the angle of incidence.

The second part of this law therefore requires that

$$\frac{\sin \phi}{\sin \phi'} = \text{const.} \quad (1b)$$

If on the left side of the boundary in Fig. 1*B* there exists a vacuum (or for practical purposes air), the value of the constant in Eq. 1*b* is called the *index of refraction* n of the medium on the right. By experimental measurements of the angles ϕ and ϕ' one can determine the values of n for various transparent substances. Then, in the refraction at a boundary between two such substances having indices of refraction n and n' , Snell's law may be written in the symmetrical form

$$n \sin \phi = n' \sin \phi' \quad (1c)$$

Wherever it is feasible we shall use *unprimed* symbols to refer to the first medium and *primed* ones for the second. The ratio n'/n is often called the *relative index* of the second medium with respect to the first. The constant ratio of the sines in Eq. 1*b* equals this relative index. When the angle of incidence is fairly small, Eq. 1*c* shows that the angle of refraction will also be small. Under these circumstances a very good approximation is obtained by setting the sines equal to the angles themselves, so we obtain

$$\frac{\phi}{\phi'} = \frac{n'}{n} \quad \text{FOR SMALL ANGLES} \quad (1d)$$

1.3. Principle of Reversibility. The symmetry of Eqs. 1*a* and 1*c* with respect to the primed and unprimed symbols shows at once that *if a reflected or refracted ray be reversed in direction, it will retrace its original path*. For a given pair of media with indices n and n' any one value of ϕ is correlated with a corresponding value of ϕ' . This will be equally true when the ray is reversed and ϕ' becomes the angle of incidence in the medium of index n' ; the angle of refraction will then be ϕ . Since the reversibility holds at each reflecting or refracting surface, it holds also for even the most complicated light paths. This useful principle has more

* Willebrord Snell (1591–1626) of the University of Leyden, Holland. He announced what is essentially this law in an unpublished paper in 1621. His geometrical construction required that the ratios of the cosecants of ϕ' and ϕ be constant. Descartes was the first to use the ratio of the sines, and the law is generally known as Descartes' law in France. (See footnote, p. 9.)

than a purely geometrical foundation, and it will be shown later that it follows from the application to wave motion of a corresponding principle in mechanics.

1.4. Optical Path. In order to state a more general principle which will include both the law of reflection and that of refraction, it is convenient to have the definition of a quantity called the *optical path*. When light travels a distance d in a medium of refractive index n the optical path is the product nd . The physical interpretation of n , to be given later, shows that the optical path represents the distance in vacuum that the light would traverse in the same time that it goes the distance d in the medium. When there are several segments d_1, d_2, \dots of the light path in substances having different indices n_1, n_2, \dots , the optical path is found as follows:

$$\text{Optical path} = [d] = n_1 d_1 + n_2 d_2 + \dots = \sum n_i d_i \quad (1c)$$

For example let L in Fig. 1C represent a lens of refractive index n'

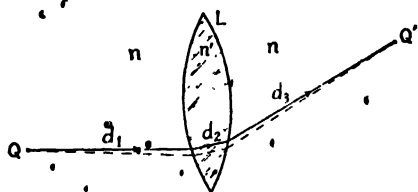


FIG. 1C. Illustrating optical path and Fermat's principle.

immersed in some liquid of index n . The optical path between two points Q and Q' on a ray becomes, in this case,

$$[d] = nd_1 + n'd_2 + nd_3.$$

Here Q and Q' need not necessarily represent points on the object and image; they are merely *any* two chosen points on an actual ray.

One may also define an optical path in a medium of continuously varying refractive index by replacing the summation by an integral. The paths of the rays are then curved, and the law of refraction loses its meaning. We shall now consider a principle which is applicable for any type of variation of n and hence contains within it the laws of reflection and refraction as well.

1.5. Fermat's* Principle. A correct and complete statement of this principle is seldom found in textbooks, because the tendency is to cite it in Fermat's original form, which was incomplete. Using the concept of optical path, the principle should read

* Pierre Fermat (1608 1665). French mathematician, ranked by some as the discoverer of differential calculus. The justification of his principle given by Fermat was that "nature is economical," but he was unaware of circumstances where exactly the reverse is true.

The path taken by a light ray in going from one point to another through any set of media is such as to render its optical path equal, in the first approximation, to other paths closely adjacent to the actual one.

The "other paths" must be possible ones in the sense that they may only undergo deviations where there are reflecting or refracting surfaces. Now Fermat's principle will hold for a ray whose optical path is a *minimum* with respect to adjacent hypothetical paths. Fermat himself stated that the time required by the light to traverse the path is a minimum, and the optical path is a measure of this time. But there are plenty of cases in which the optical path is a *maximum*, or else neither a maximum nor a minimum but merely *stationary* (at a point of inflection) at the position of the true ray. A special case of the latter type occurs where the path is *constant*, as it would be in Fig. 1C, for example, if Q and Q' were related as object and image. Assuming the lens to be perfect, each ray leaving Q would reach Q' having traversed exactly the same optical path. The essential condition involved in Fermat's principle is that any slight variation of the actual path, for example to the broken line of Fig. 1C, must at most cause only a second-order variation in the optical path. The term "stationary" expresses this condition, and also includes the possibilities of having the optical path a maximum or a minimum.

It is easily shown that both the laws of reflection and refraction follow at once from this principle. Figure 1D(a), which represents the refraction

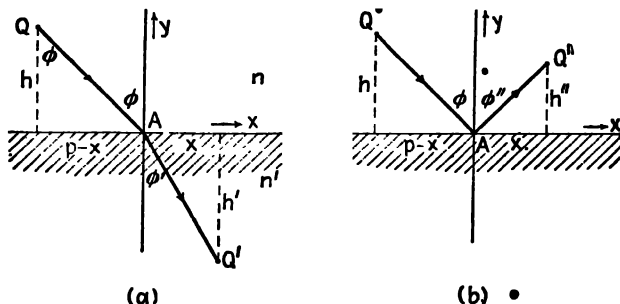


FIG. 1D. Path of a ray (a) refracted and (b) reflected at a plane surface.

tion of a ray at a plane surface, may be used to prove the law of refraction, Eq. 1c. The length of the optical path between a point Q in the upper medium (index n) and another point Q' in the lower medium (index n'), passing by any point A on the surface, is

$$[d] = nQA + n'AQ'$$

Calling the perpendicular distances to the surface h and h' and the total length of the x axis intercepted by these perpendiculars p , we have

$$[d] = n[h^2 + (p - x)^2]^{\frac{1}{2}} + n'[(h')^2 + x^2]^{\frac{1}{2}}$$

According to Fermat's principle, $[d]$ must be a maximum or minimum (or in general stationary) for the actual path. The mathematical statement of this requirement is that the first derivative with respect to x must vanish, or that

$$\frac{d[d]}{dx} = \frac{\frac{1}{2}n}{[h^2 + (p - x)^2]^{\frac{1}{2}}} (-2p + 2x) + \frac{\frac{1}{2}n'}{[(h')^2 + x^2]^{\frac{1}{2}}} (2x) = 0$$

This gives

$$n \frac{p - x}{[h^2 + (p - x)^2]^{\frac{1}{2}}} = n' \frac{x}{[(h')^2 + x^2]^{\frac{1}{2}}}$$

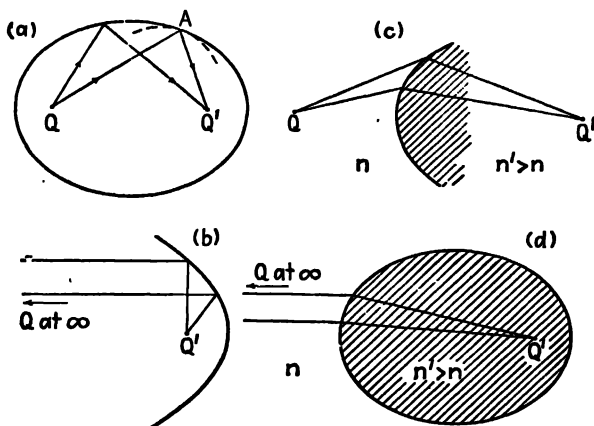


FIG. 1E. Aplanatic surfaces (a) and (b) for reflection, (c) and (d) for refraction.

By reference to Fig. 1D(a) it will be seen that the multipliers of n and n' are just the sines of the corresponding angles, so that we have now proved Eq. 1c, namely

$$n \sin \phi = n' \sin \phi' \quad (1c)$$

Application of the same method to part (b) of the figure will give $\sin \phi = \sin \phi''$, or $\phi = \phi''$, thus proving the law of reflection.

1.6. Aplanatic Surfaces. A surface of such a shape that it brings together all the rays emanating from a point source Q at some other point Q' is called an *aplanatic surface*. For any given aplanatic surface, Q and Q' are called the *aplanatic points*. Since all rays drawn from Q

to Q' by reflection or refraction at an aplanatic surface are possible rays, Fermat's principle tells us that the optical lengths of all these paths must be equal.

The aplanatic surface for reflection is in general an ellipsoid of revolution or *spheroid* having Q and Q' as foci. Figure 1E(a) illustrates this case. It is a well-known property of the ellipse that the distance $QA + AQ'$ is constant for any position of A , and also that QA and AQ' make equal angles with the normal drawn at A (law of reflection). Incidentally, a simple example of a case where Fermat's principle requires the path to be a *maximum* may be seen by comparing the surface given by the broken line in Fig. 1E(a) with that of the spheroid. This surface is more curved than the spheroid but is tangent to it at A . It is clear that the path of a ray reflected at A will be greater than other paths drawn to adjacent points on the broken curve. When one of the points, either Q or Q' , is at an infinite distance, as in Fig. 1E(b), the aplanatic surface for reflection becomes a paraboloid of revolution. This is the shape used in astronomical telescope mirrors and in searchlight reflectors.

The aplanatic surfaces for refraction are somewhat more complicated, since they are defined by the equation

$$n(QA) + n'(AQ') = \text{const.}$$

These surfaces were first investigated by Descartes* and are called *cartesian ovals* in honor of him. An example is shown in Fig. 1E(c). In the case where one of the aplanatic points is at infinity they become conic sections, as for example in the spheroid of Fig. 1E(d). Aplanatic refracting surfaces are of little practical use in optics because they give a perfect image for only one point at one distance. An exception to this statement is found, however, in the use of the aplanatic points of a sphere in the oil-immersion microscope (Sec. 9.6).

1.7. Pencils, Beams, and Bundles of Rays. A narrow cone of rays coming from a point source, or from some one point of a broad source, is called a *pencil* of rays. Specifically it is a *homocentric pencil*, since all the rays when projected backward pass through a common center. A homocentric pencil will remain homocentric after reflection or refraction only if the surface is an aplanatic surface [Fig. 1F(a)]. Any other form of surface will render the reflected or refracted pencil nonhomocentric, as in (b) of Fig. 1F. Hence it is rather rare that a refracted pencil remains truly homocentric.

The summation of all the pencils coming from the various points on a broad source is called a *beam* of light. Thus, in Fig. 1F(c), to represent

* René Descartes (1596-1650). French mathematician and philosopher. He was an early proponent of the use of momentum and of the conservation of momentum.

the complete beam one would have to draw divergent pencils from each point on the source S . It is worth noting that even though the individual pencils after passing through the lens L may consist of parallel rays, the resulting beam does not constitute truly *parallel light*, since there are rays in it which make considerable angles with each other. The light will be more nearly parallel the smaller the size of the source relative to its distance from the lens.

The term *bundle* of rays is used to designate the pencil which, starting from some point on the light source, traverses an optical instrument from one end to the other. Somewhere within any instrument the width of

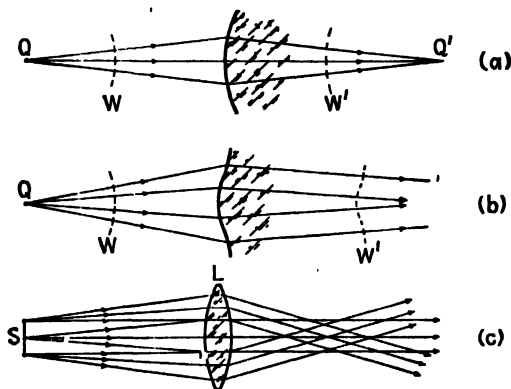


FIG. 1F. Illustrating (a) homocentric pencils; (b) refracted pencil not homocentric; (c) refracted beam consisting of parallel homocentric pencils.

the system of rays is limited by an aperture which determines the effective rays to be included in the bundle. The bundle is therefore defined in terms of a particular optical instrument. The subject of apertures and stops will be discussed in detail in Chap. 7.

1.8. Wave Fronts. When light travels in an isotropic medium like glass or air, in which the velocity of the light does not vary with direction, any surface which is everywhere perpendicular to the rays in a pencil is called a *wave front*. In certain crystals this situation does not hold (see Sec. 25.2), but we shall assume for the present that we are dealing only with isotropic media. The wave fronts associated with homocentric pencils are spherical, as shown at W and W' in Fig. 1F(a) and at W in Fig. 1F(b). Consideration of wave fronts is useful in certain parts of geometrical optics, and it is not necessary here to know their physical significance. For the present they may be considered as *surfaces drawn perpendicular to the rays* in a pencil. The significance of the term "wave front" will be explained when we come to the subject of wave motion in Chap. 11.

1.9. Malus' Theorem.* This theorem states that when a pencil of rays traverses two or more media, and hence has been refracted one or more times, a new wave front may always be found in one of the subsequent media by measuring off equal optical paths along all rays, starting from a wave front in the first medium. In other words *the optical path between any two wave fronts is the same for any ray*. It is not difficult to derive this theorem from Fermat's principle, and the two are in fact entirely equivalent. Thus in Fig. 1G, which is Fig. 1F(b) shown in greater detail, let the surface W' be drawn so that it cuts each ray at the point where its optical path from W is the same. Then Malus' theorem states that this is a wave front, *i.e.*, is perpendicular to all rays. To prove this we have drawn two rays Q_1A_1 and Q_2A_2 , which are close together and each perpendicular to the surface W . By construction they have equal optical paths to W' , so that

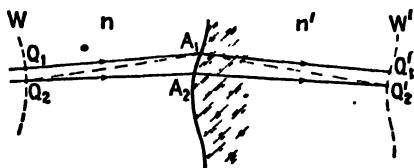


FIG. 1G. Illustrating Malus' theorem.

$$[d] = nQ_1A_1 + n'A_1Q'_1 = nQ_2A_2 + n'A_2Q'_2 \quad (1f)$$

Now we apply Fermat's principle to the true path $Q_2A_2Q'_2$ and the closely adjacent path $Q_2A_1Q'_2$ (broken lines). It is required that

$$nQ_2A_2 + n'A_2Q'_2 \cong nQ_2A_1 + n'A_1Q'_2 \quad (1g)$$

where the symbol \cong denotes equality when second-order terms are neglected. To this degree of approximation we then have, from Eqs. 1f and 1g,

$$nQ_2A_1 + n'A_1Q'_2 \cong nQ_1A_1 + n'A_1Q'_1 \quad (1h)$$

Since by construction $Q_1A_1 \perp Q_1Q_2$, we know that $Q_1A_1 \cong Q_2A_1$, and hence that

$$nQ_1A_1 \cong nQ_2A_1 \quad (1i)$$

Upon subtracting Eq. 1i from Eq. 1h, there results

$$n'A_1Q'_2 \cong n'A_1Q'_1$$

so that finally $A_1Q'_2 \cong A_1Q'_1$, and $Q'_1Q'_2 \perp A_1Q'_1$. The surface W' is therefore everywhere perpendicular to the rays and by definition consti-

* Étienne Malus (1775–1812). French army engineer. His most celebrated discovery was that of polarization by reflection (Chap. 24), which he observed by accident when looking through a calcite crystal at the light reflected from the windows of the Luxembourg palace.

tutes a wave front. This proof can be readily extended to the case where several surfaces separating different media are involved.

1.10. Huygens'* Construction. The theorem of Malus may be used to find the refracted wave front, and hence the refracted rays, through a geometrical construction due to Huygens. In Fig. 1H(a) suppose it is

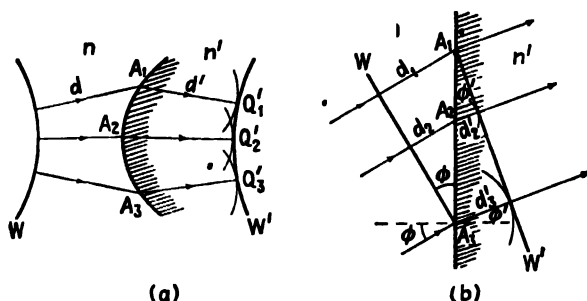


FIG. 1H. Huygens' construction (a) when the wave front and the surface have arbitrary shapes; (b) when a plane wave strikes a plane boundary.

required to find the wave front W' which is separated by the constant optical[†] path

$$[d] = nd + n'd' \quad (1j)$$

from the original wave front W . Evidently d and d' will be different for each ray, but if the value of $[d]$ is specified we can measure d for each, and compute the corresponding values of d' from Eq. 1j. If we then draw circles about the appropriate points such as A_1 , A_2 , and A_3 using these values of d' as radii, we see that the new wave front W' must be tangent to all the circles, and furthermore that the refracted rays must be drawn to the points of tangency Q'_1 , Q'_2 , and Q'_3 . Only in this way can the wave front be always perpendicular to the ray and at the same time be consistent with Malus' theorem.

As an example of the usefulness of Huygens' construction, we may give here the rather familiar proof of the law of refraction. Referring to Fig. 1H(b), the constant optical path between W and W' for any ray such as the one incident at A_2 is seen to be

$$[d] = nd_2 + n'd'_2 = n(A_2A_3) \sin \phi + n'd'_2$$

* Christian Huygens (1629-1695). Famous Dutch scientist and contemporary of Isaac Newton. Huygens' principal contributions were in the wave theory of light (Chap. 13), but he also made valuable discoveries in dynamics, mathematics, and astronomy. It is said that Huygens got his first ideas about the propagation of waves by watching the ripples on a Dutch canal.

But this must also represent the optical path difference for the ray whose path d_1 lies entirely in the first medium, namely

$$[d] = nd_1 = n(A_1A_2) \sin \phi$$

Equating the two values of $[d]$ and substituting $A_1A_3 - A_1A_2$ for A_2A_3 , one easily finds that *

$$d'_2 = \frac{n}{n'} (A_1A_2) \sin \phi \quad (1k)$$

The fact that d'_2 is directly proportional to the distance A_1A_2 shows that the refracted wave remains plane. Furthermore $d'_2/(A_1A_2) = \sin \phi'$, and substituting this value in Eq. 1k, we again obtain Snell's law

$$n' \sin \phi' = n \sin \phi$$

Once it is known that W' is plane, it is only necessary to construct the circle of radius d'_2 and draw the tangent that passes through A_1 .

1.11. Color Dispersion. It is well known to those who have studied elementary physics that refraction causes a separation of white light into its component colors. Thus, as is shown in Fig. 1I, the incident ray of white light gives rise to refracted rays of different colors (really a continuous *spectrum*) each of which has a different value of ϕ' . By Eq. 1c the value of n' must therefore vary with color. It is customary in the exact specification of indices of refraction to use the particular colors corresponding to certain dark lines in the spectrum of the sun. A table of these so-called Fraunhofer* lines, which are designated by the letters A, B, . . . , starting at the extreme red end, is given later in Sec. 21.10. The ones most commonly used are those in Fig. 1I.

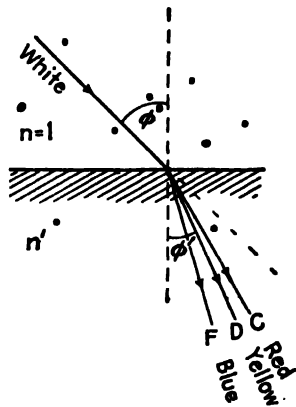


FIG. 1I. Refraction of a ray of white light.

The angular divergence of rays F and C is a measure of the *dispersion* produced, and has been greatly exaggerated in the figure relative to the average *deviation* of the spectrum, which is

* Joseph Fraunhofer (1787–1826). Son of a poor Bavarian glazier, Fraunhofer learned glass grinding, and entered the field of optics from the practical side. His rare experimental skill enabled him to produce much better spectra than those of his predecessors and led to his study of the solar lines with which his name is now associated. Fraunhofer was one of the first to produce diffraction gratings (Chap. 17).

measured by the angle through which ray D is bent. To take a typical case of crown glass, the refractive indices as given in Table 23I are,

$$n_F = 1.5330, \quad n_D = 1.5270, \quad n_C = 1.5244$$

Now it is readily shown from Eq. 1d that for a given small angle ϕ the dispersion of the F and C rays ($\phi'_F - \phi'_C$) is proportional to $n_F - n_C = 0.0086$, while the deviation of the D ray ($\phi - \phi'_D$) depends on $n_D - 1 = 0.5270$ and is thus more than sixty times as great. The ratio of these two quantities varies greatly for different kinds of glass and is an important characteristic of any optical substance. It is called the *dispersive power* and is defined by the equation

$$1 \quad \frac{n_F - n_C}{n_D - 1} \quad (1l)$$

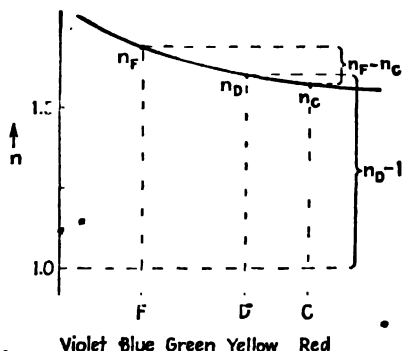


FIG. 1J. Variation of refractive index with color.

The reciprocal of the dispersive power, designated by the Greek letter ν , lies between 30 and 60 for most optical glasses.

Figure 1J illustrates schematically the type of variation of n with color that is usually encountered for optical materials. The numerator

of Eq. 1l, which is a measure of the dispersion, is determined by the difference in the index at two points near the ends of the spectrum. The denominator, which measures the average deviation, represents the magnitude in excess of unity of an intermediate index of refraction.

1.12. Diffusion and Absorption of Light. These are processes which weaken or destroy light rays. The laws of reflection and refraction apply to surfaces smooth enough to yield *regularly reflected and transmitted* pencils. In nature the majority of surfaces are not of this kind but contain small particles or fibers which scatter the light in all directions as illustrated in Fig. 1K(a) and (b). We then speak of *diffuse reflection* and *diffuse transmission*. The cause of this phenomenon and the laws governing it will be discussed in later chapters. For the present we may assume that it will be negligible for the surfaces of optical parts like lenses.

Absorption of a light beam, i.e., the conversion of the energy of the light into heat, will always occur to a greater or less extent in reflection and transmission. When light is reflected from a highly polished metal

it penetrates a minute distance into the metal, and a certain fraction of the light does not appear in the reflected beam because of absorption [Fig. 1K(c)]. Similarly when light passes through any medium, even a so-called "transparent" one, some of the light always fails to be transmitted [Fig. 1K(d)]. The absorption is not usually equally strong for different colors. For example, large thicknesses of glass or of water have less absorption for green light. Very strong absorption, as occurs for instance in a metal, is always accompanied by strong reflection.

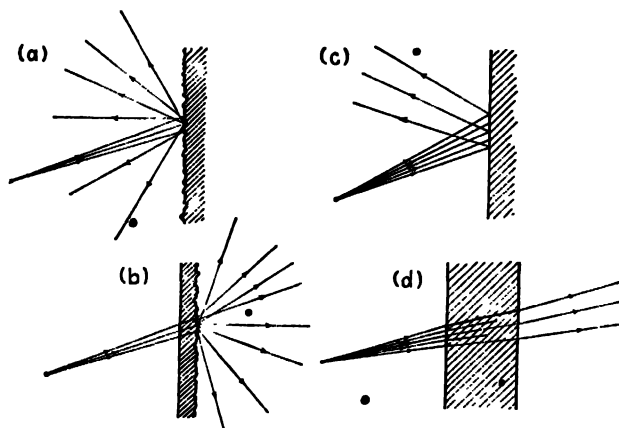


FIG. 1K. Schematic representation of (a) diffuse reflection; (b) diffuse transmission; (c) absorption upon reflection; (d) absorption upon transmission.

Absorption and diffusion are processes which affect the intensity of light and the brightness and contrast of optical images. We shall first be concerned, however, with the position, size, and shape of these images, which are determined by the basic laws of geometrical optics given in the preceding sections, and shall return to questions of intensity in Chap. 7.

Problems

1. In the experiment of Fig. 1A, if the source and observing screen are each 1 m from the hole H and the latter is 0.5 mm in diameter, diffraction will spread the light on the screen into a circular patch 2 mm in diameter. What is the maximum permissible diameter of the "point" source in order that spreading due to the finite size of this source shall not exceed one-tenth of that due to diffraction?

2. The angle of incidence of a ray on the surface of water ($n' = 1.3330$) is 10° . What percentage error in the angle of refraction is made by assuming that in Snell's law the sines may be replaced by the angles as in Eq. 1d?

3. An approximate law of refraction was given by Kepler in the form $\phi = \phi' / (1 - k \sec \phi')$, where $k = (n' - 1)/n'$, n' being the relative index. Plot curves on the same set of coordinates showing the variation of ϕ' with ϕ according to this formula and according to the true Eq. 1c.

4. Parallel light falls on a 60° glass prism of refractive index 1.50, the angle of incidence on the first surface being 30° . Using Huygens' construction and Malus' theorem, draw the position of (a) a wave front in the prism, and (b) another wave front after emergence from the prism.

5. Assuming the length of the base of the prism in Prob. 4 to be 8 cm, calculate the optical path in centimeters between the first and last contacts of the wave front with the prism when the angle of incidence is 30° .

6. A long pipe is closed at either end with a glass plate 5 mm thick (refractive index of glass = 1.57) and is evacuated. If the separation between the outer surfaces of the two glass windows is exactly 10 m, what is the optical path between these surfaces? By how much is it changed if the pipe is filled with air at one atmosphere pressure (refractive index of air = 1.000277)?

7. A ray passes perpendicularly through a plane-parallel glass plate 8 mm thick and having an index of refraction of 1.620. If the plate is turned through an angle of 5° about an axis perpendicular to the ray, what is the increase in the optical path?

8. Draw a ray between two points Q and Q' , both of which lie on the axis of a thin lens. The ray passes through a point A in the lens, this point not being on the axis. Extend AQ' to some point Q'' beyond the focus Q' . By considering the change in the path QAQ'' when the position of A is changed slightly, find whether the actual path QAQ'' is a maximum, a minimum, or otherwise stationary.

9. A glass sphere of radius 5 cm has $n = 1.667$. Let a straight ray be drawn from a point on the surface of this sphere, through its center and to a point 25 cm away, i.e., 15 cm outside the surface of the sphere on the other side. Find, by graphical construction and measurement of the paths, whether this ray has a path which is a maximum or a minimum. Draw in roughly the form of the aplanatic surface for these two points, making it tangent to the spherical surface.

10. (a) Calculate the dispersive powers and ν -values for water ($n_F = 1.33714$, $n_D = 1.33300$, and $n_C = 1.33115$), and also for the crown glass mentioned in Sec. 1.11. (b) Sunlight is incident on a surface of this crown glass immersed in water, its rays making an angle of 60° with the normal. Find the angular separation of the C and F rays in the glass.

11. Two plane mirrors are inclined at a fixed angle α , and can be rotated about their line of intersection as an axis. Using the law of reflection, show that any ray whose plane of incidence is perpendicular to this axis is deviated in the two reflections by an angle which is independent of the rotation of the mirrors. Express this deviation in terms of α .

12. Show from the equation of a parabola, $y = ax^2$, that a paraboloid constitutes the aplanatic surface for a point at $y = +\infty$ and the focus of the parabola. This amounts to showing that the path between any incident plane wave front and the focus is constant.

CHAPTER 2

PLANE SURFACES

The behavior of a pencil of light upon reflection or refraction at a plane surface is of basic importance in geometrical optics. Its study will reveal several of the features that will later have to be considered in the more difficult case of a curved surface. Plane surfaces often occur in nature, for example as the cleavage surfaces of crystals or as the surfaces of liquids. Artificial plane surfaces are used in optical instruments to bring about deviations or lateral displacements of rays as well as to break light into its colors. The most important devices of this type are prisms, but before taking up this case of two surfaces inclined to each other, we must examine rather thoroughly what happens at a single plane surface.

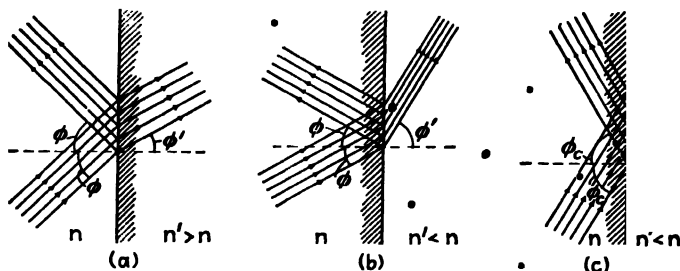


FIG. 2A. Reflection and refraction of a parallel beam. (a) External reflection; (b) internal reflection at an angle smaller than the critical angle; (c) total internal reflection at the critical angle.

2.1. Parallel Beam. In a beam or pencil of parallel light, each ray meets the surface traveling in the same direction. Therefore any one ray may be taken as representative of all the others. The parallel beam remains parallel after reflection or refraction at a plane surface, as is shown in Fig. 2A(a). Refraction causes a change in width of the beam which is easily seen to be in the ratio $\cos \phi' / \cos \phi$, whereas the reflected beam remains of the same width. This will prove to be important for intensity considerations (Sec. 28.1). There is also dispersion of the refracted beam (Sec. 1.11) but not of the reflected one. The incident pencil, which is homocentric with its center at infinity, remains homocentric in both cases. One may say that a plane surface is an aplanatic surface (Sec. 1.6) with both its apianatic points at infinity.

Reflection at a surface where n increases, as in Fig. 2A(a), is called *external reflection*. It is also frequently termed *rare-to-dense* reflection because the relative magnitudes of n correspond roughly (though not exactly) to those of the actual densities of materials. In Fig. 2A(b) is shown a case of *internal reflection* or *dense-to-rare* reflection. In this particular case the refracted beam is narrow because ϕ' is close to 90° .

2.2. Total Reflection. The particular value of ϕ for which $\phi' = 90^\circ$ is called the *critical angle* ϕ_c . At values of ϕ greater than this there can be no refracted beam, and the light therefore undergoes *total reflection*. An equation giving the critical angle is obtained by substituting $\phi' = 90^\circ$, or $\sin \phi' = 1$, in the law of refraction, Eq. 1c:

$$n \sin \phi_c = n' \times 1$$

so that

$$\sin \phi_c = \frac{n'}{n} \quad (2a)$$

a quantity always less than unity. Total reflection* is really total in the sense that no energy is lost upon reflection. In any device intended to

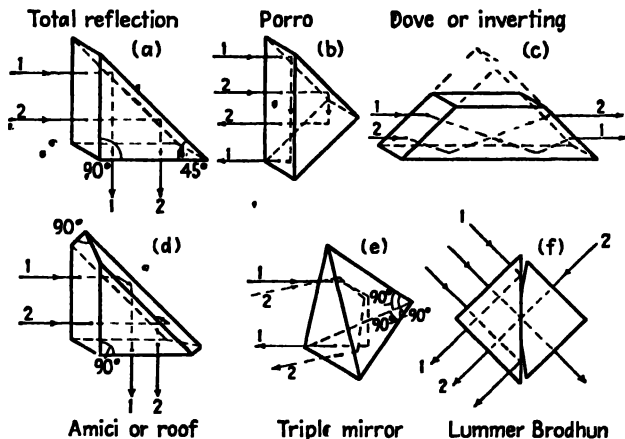


FIG. 2B. Reflecting prisms utilizing the principle of total reflection.

utilize this property there will, however, be small losses due to absorption in the medium and to reflections at the surfaces where the light enters and leaves the medium. The commonest device of this kind is the *total reflection prism*, which is a glass prism with two angles of 45° and one of 90° . As shown in Fig. 2B(a), the light usually enters perpendicular to one of the shorter faces, is totally reflected from the hypotenuse, and leaves at right angles to the other short face. This deviates the rays through a right angle. Such a prism may also be used in two other

ways which are illustrated in (b) and (c) of the figure. The Dove prism (c) interchanges the two rays, and if the prism is rotated about the direction of the light, they rotate around each other with twice the angular velocity of the prism.

Many other forms of prisms which use total reflection have been devised for special purposes. Two common ones are illustrated in Fig. 2B(d) and (e). The roof prism accomplishes the same purpose as the total reflection prism (a) except that it introduces an extra inversion. The triple mirror (e) is made by cutting off the corner of a cube by a plane which makes equal angles with the three faces intersecting at that corner. It has the useful property that any ray striking it will, after being internally reflected at each of the three faces, be sent back parallel to its original direction. The Lummer-Brodhun "cube" shown in (f) is used in photometry to compare the illumination of two surfaces, one of which is viewed by rays (2) coming directly through the circular region where the prisms are in contact, the other by rays (1) which are totally reflected in the area around this region.

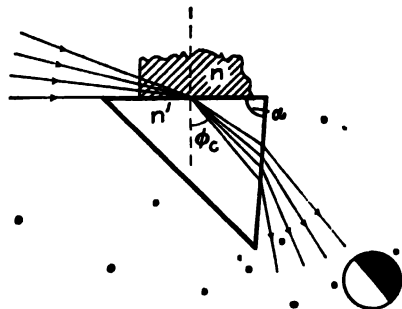


FIG. 2C. Principle of a refractometer.

Since, in the examples shown, the angle of incidence is always 45° , it is essential that this shall exceed the critical angle in order that the reflection be total. Supposing the outer medium to be air ($n' = 1$), this requirement sets a lower limit on the value of the index n of the prism. By Eq. 2a we must have

$$\frac{n'}{n} = \frac{1}{n} \geq \sin 45^\circ$$

so that $n \geq \sqrt{2} = 1.414$. This condition always holds for glass and is even fulfilled for optical materials having low refractive indices such as lucite ($n = 1.49$) and fused quartz ($n = 1.46$).

The principle of most accurate *refractometers* (instruments for the determination of refractive index) is based on the measurement of the critical angle ϕ_c . In both the Pulfrich and Abbe types a convergent beam strikes the surface between the unknown sample, of index n , and a prism of known index n' . Now n' is greater than n , so the two must be interchanged in Eq. 2a. The beam is so oriented that some of its rays just graze the surface (Fig. 2C), so that one observes in the transmitted light

a sharp boundary between light and dark. Measurement of the angle at which this boundary occurs allows one to compute the value of ϕ , and hence of n . There are important precautions that must be observed if the results are to be at all accurate.*

2.3. Reflection of a Divergent Pencil. When a homocentric pencil of any degree of divergence is reflected at a plane surface, it remains homo-

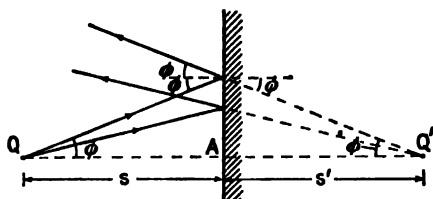


FIG. 2D. Reflection of a divergent pencil.

centric. All rays originating from a point Q (Fig. 2D) will after reflection appear to come from another point Q' symmetrically placed behind the mirror. The proof of this proposition follows at once from the application of the law of reflection (Eq. 1a), according to which all the angles labeled ϕ in the figure must be equal. Under these conditions the distances QA and AQ' along the line QAQ' drawn perpendicular to the surface must be equal: i.e.,

$$s' = s \quad (2b)$$

The point Q' is said to be a *virtual image* of Q , since when the eye receives the reflected rays they appear to come from a source at Q' but do not actually pass through Q' as would be the case if it were a *real image*. In order to produce a real image a surface other than a plane one is required.

2.4. Refraction of a Divergent Pencil. As was stated in Sec. 1.7, the refraction of a homocentric pencil will in general render it nonhomocentric. Referring to Fig. 2E, let us find the position of the point Q' where the lower refracted ray, when produced backward, crosses the perpendicular to the surface drawn through Q . Let $QA = s$, $Q'A = s'$, and $AB = h$. Then

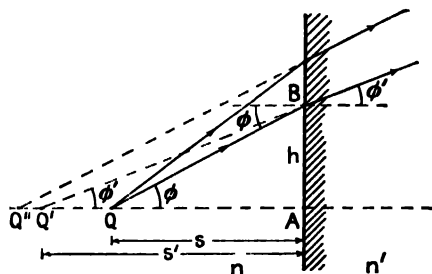


FIG. 2E. Refraction of a divergent pencil.

$$h = s \tan \phi = s' \tan \phi'$$

* For a valuable description of this and other methods of determining indices of refraction see A. C. Hardy and F. H. Perrin, "Principles of Optics," 1st ed., pp. 359-364, McGraw-Hill Book Company, Inc., New York,

so that

$$s' = s \frac{\tan \phi}{\tan \phi'} = s \frac{\sin \phi \cos \phi'}{\sin \phi' \cos \phi}$$

Now according to the law of refraction, Eq. 1c, the ratio

$$\frac{\sin \phi}{\sin \phi'} = \frac{n'}{n} = \text{constant.}$$

We therefore have

$$s' = s \frac{n' \cos \phi'}{n \cos \phi} \quad (2c)$$

The ratio of the cosines is not constant. Instead, starting at the value unity for small ϕ , it increases slowly at first, then more rapidly. As a consequence the projected rays do not intersect at any single point such as Q' . Furthermore they do not all intersect at any other point in space; i.e., the refracted pencil is nonhomocentric. For a further description of the characteristics of this pencil see Sec. 2.6 below.

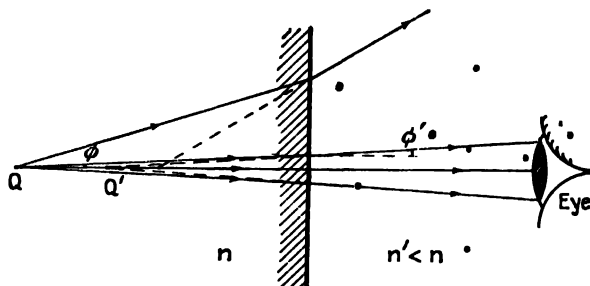


FIG. 2F. Image by refraction of paraxial rays by a plane surface.

2.5. Images Formed by Paraxial Rays. It is well known that when one looks at objects through the plane surface of a refracting medium, as for example in an aquarium, the objects are seen clearly. Actually one is seeing virtual images which are not in the true position of the objects. When one looks perpendicularly into water they appear closer to the surface in about the ratio $3/4$, which is the ratio n'/n , since $n' = 1$ for air and $n = 1.33 \cong 4/3$ for water. This observation is readily understood when one considers that the rays entering the pupil of the eye will in this case make extremely small angles with the normal to the surface, as shown in Fig. 2F. Therefore both cosines in Eq. 2c are nearly equal to unity, and their ratio is even more nearly so. Hence, as long as the rays are restricted to ones that make very small angles with the normal

to the refracting surface, a good virtual image is formed at the distance s' given by

$$s' = \frac{n'}{n} s \quad \text{PARAXIAL RAYS} \quad (2d)$$

Rays for which the angles are small enough so that we may set the cosines equal to unity, and the sines equal to the angles, are called *paraxial rays*.

2.6. Astigmatic Pencil. When a narrow incident pencil makes a large angle with the normal, the refracted pencil departs radically from the homocentric character obtained above. It is said to be *astigmatic*. To explain this term we refer first to Fig. 2G(a), in which the extreme rays of the refracted pencil when projected backward are seen to cross the normal at different points as is required by Eq. 2c. Let their intersection

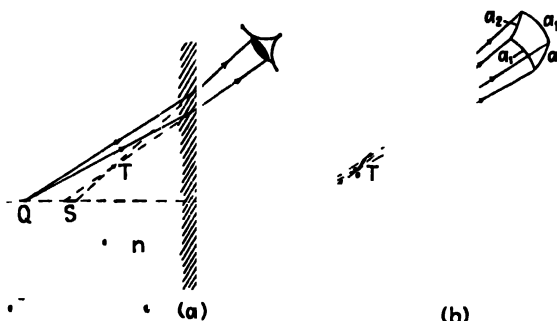


FIG. 2G. Astigmatism of an oblique pencil.

lie at T . This is not however a virtual *point* image of Q as would appear from the diagram. For since the angle of refraction is constant at a given angle of incidence, the shape of the pencil in space is obtained by rotating the figure slightly about the line QA as an axis. T then traces out a short line perpendicular to the figure. When all the rays are projected backward, they are found to cross somewhere along the line T and again along a second line S coinciding with the normal. The eye sees a somewhat blurred image of Q located somewhere between these two focal lines T and S .

Consideration of the form of the refracted wave front will help to understand the nature of astigmatism. In Fig. 2G(b) the refracted pencil is magnified and drawn in perspective. A rectangular segment of the wave front is also shown, the sides a_1 and a_2 being respectively in and perpendicular to the plane of the figure. From the above description of the rays it will be seen that the wave front has *different curvatures in these two mutually perpendicular directions*. The center of curvature of the arcs a_1 lie on T and of the arcs a_2 on S . This is the characteristic

feature of an astigmatic wave front and, as we shall see, is of common occurrence. Finally it should be mentioned that, for a very wide pencil occupying a cone centered on QA , the summation of all astigmatic pencils gives rise to a virtual focus in the shape of a *caustic*, and the effect is known as *spherical aberration*. Detailed consideration of this phenomenon will be postponed until we come to the more easily visualized case of a real focus (Sec. 6.9).

2.7. Plane-parallel Plate. When a single ray traverses a glass plate with plane surfaces that are parallel to each other, it emerges parallel to its original direction but with a lateral displacement d which increases

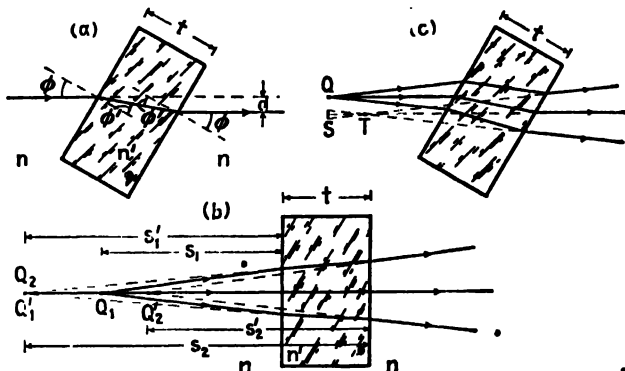


FIG. 2H. Refraction by a plane-parallel plate.

with the angle of incidence ϕ . Using the notation shown in Fig. 2H(a), we may apply the law of refraction and some simple trigonometry to show that the displacement is given by

$$d = t \sin \phi \left(1 - \frac{n \cos \phi}{n' \cos \phi'} \right) \quad (2e)$$

For small angles, d is nearly proportional to $\sin \phi$, but the ratio of the cosines soon becomes appreciably less than 1, causing a somewhat more rapid rate of increase.

If we now consider a divergent pencil to be incident on such a plate, the different rays of the pencil are not all incident at exactly the same angle ϕ , and therefore they undergo slightly different lateral shifts. For paraxial rays this yields a point image which is shifted toward the plate by a distance $t \left(1 - \frac{n}{n'} \right)$.

This result may easily be obtained by applying Eq. 2d successively for the two surfaces, considering the image due to the first surface to be the

object for the second. Referring to Fig. 2H(b), let Q_1 be the point source at a distance s_1 from the first surface. For the distance to the image Q'_1 formed by this surface, Eq. 2d yields

$$s'_1 = \frac{n'}{n} s_1$$

The rays strike the second surface as though they come from Q'_1 , which may therefore be also considered as a point object Q_2 for this surface. It is at the distance $s_2 = s'_1 + t$ from this surface, so that one has for the final image distance

$$s'_2 = \frac{n}{n'} s_2 = \frac{n}{n'} (s'_1 + t) = \frac{n}{n'} \left(\frac{n'}{n} s_1 + t \right) = s_1 + \frac{n}{n'} t$$

Therefore the displacement of the image is

$$Q_1 Q'_2 = s_1 - (s'_2 - t) = t \left(1 - \frac{n}{n'} \right)$$

as stated above.

When the plate is turned through an appreciable angle as in part (c) of Fig. 2H, the emergent pencil becomes astigmatic, because the lateral

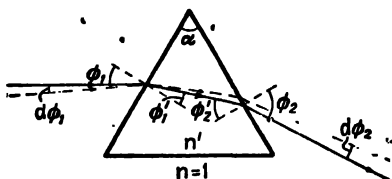


FIG. 2I. Angular magnification by a prism (monochromatic light).

displacements of the rays are such that their projections no longer pass even approximately through a point. This leads, as in the case of a single surface, to the formation of two virtual focal lines T' and S .

2.8. Refraction by a Prism. In a prism the two surfaces are inclined at some angle α so that the deviation produced by the first surface is not annulled by the second but is further increased. The dispersion of color (Sec. 1.11) is also increased, and this is usually the main function of a prism. First let us consider, however, the geometrical optics of the prism for light of a single color, i.e., for *monochromatic light* such as is obtained from a sodium arc. The solid ray in Fig. 2I shows the path of a ray incident on the first surface at the angle ϕ_1 . We assume that the prism is in air or else that n' represents the *relative* index of the prism with respect to its surroundings.

First it is of interest to find the change $d\phi_2$ in the angle of emergence that results from a small change $d\phi_1$ in the angle of incidence (see the

broken line in Fig. 2I). Applying the law of refraction to the first surface we have

$$\sin \phi_1 = n' \sin \phi'_1$$

To obtain the change of ϕ'_1 with ϕ_1 we differentiate, obtaining

$$\cos \phi_1 d\phi_1 = n' \cos \phi'_1 d\phi'_1 \quad (2f)$$

Similarly, for the second surface,

$$\cos \phi_2 d\phi_2 = n' \cos \phi'_2 d\phi'_2 \quad (2g)$$

Hence, dividing Eq. 2g by Eq. 2f,

$$\frac{d\phi_2}{d\phi_1} = \frac{\cos \phi_1 \cos \phi'_2 d\phi'_2}{\cos \phi_2 \cos \phi'_1 d\phi'_1} \quad (2h)$$

From the geometry of Fig. 2I one finds that $\phi'_1 + \phi'_2 = \alpha$, the *refracting angle*. Differentiation therefore gives

$$d\phi'_1 = -d\phi'_2$$

Substituting in Eq. 2h there results

$$\frac{d\phi_2}{d\phi_1} = - \frac{\cos \phi_1 \cos \phi'_2}{\cos \phi_2 \cos \phi'_1} \quad (2i)$$

Thus if one looks through a prism at a monochromatic source which subtends a small angle $d\phi_1$, its apparent width $d\phi_2$ will vary with the angle of incidence. When ϕ_1 is decreased somewhat from the value illustrated in Fig. 2I, ϕ_2 will approach 90° and the light leaves at "grazing emergence." Then $\cos \phi_2 = 0$, and the *angular magnification* $d\phi_2/d\phi_1 = \infty$. The apparent width of the source is infinite. At "grazing incidence" $\phi_1 = 90^\circ$, we have $d\phi_2/d\phi_1 = 0$. Even a wide source will look like a line.

An interesting application of the latter fact is that of looking at a mercury arc through a prism oriented for grazing incidence. One can see the mercury spectrum, and even resolve the yellow doublet, without the use of a slit or lenses. One can say that the face of the prism acts as the slit.

In the special case where the ray makes equal angles with the two faces, as in Fig. 2J(a), $\phi_1 = \phi_2$ and $\phi'_1 = \phi'_2$, so that by Eq. 2i the angular magnification is unity.

2.9. Minimum Deviation. In the above-mentioned symmetrical case, where for an equilateral prism the ray traverses it parallel to the base b

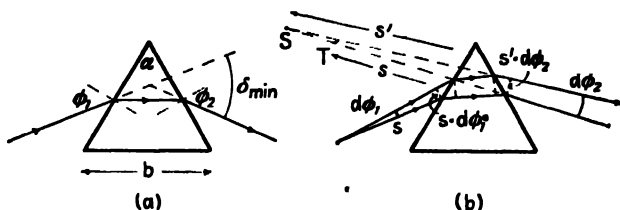


FIG. 2J. (a) Ray at minimum deviation; (b) astigmatic pencil by refraction at an angle other than that of minimum deviation.

[Fig. 2J(a)], the total deviation δ goes through a minimum. In general δ is the sum of the deviations at the two surfaces, so that

$$\delta = (\phi_1 - \phi'_1) + (\phi_2 - \phi'_2) = \phi_1 + \phi_2 - (\phi'_1 + \phi'_2)$$

or, since $\phi'_1 + \phi'_2 = \alpha$,

$$\delta = \phi_1 + \phi_2 - \alpha \quad (2j)$$

At a minimum the first derivative must vanish. Hence we have

$$\frac{d\delta}{d\phi_1} = 1 + \frac{d\phi_2}{d\phi_1} = 0$$

Accordingly, by Eq. 2i,

$$\frac{\cos \phi_1 \cos \phi'_2}{\cos \phi_2 \cos \phi'_1} = 1 \quad (2k)$$

Evidently this relation is satisfied when $\phi_2 = \phi_1$ and $\phi'_2 = \phi'_1$. The case where $\phi_2 = -\phi_1$ and hence $\phi'_2 = -\phi'_1$ is trivial, since it corresponds to a prism having $\alpha = 0$.

Measurements with a prism are usually made at the position of minimum deviation. One reason is that only here does there exist a simple relation between the refractive index and the angle of deviation. Putting $\phi_1 = \phi_2$ in Eq. 2j and noting that at minimum deviation $\alpha = 2\phi'_1 = 2\phi'_2$, there results

$$n' = \frac{\sin \phi_1}{\sin \phi'_1} = \frac{\sin \phi_2}{\sin \phi'_2} = \frac{\sin \frac{1}{2}(\delta_{min} + \alpha)}{\sin \frac{1}{2}\alpha} \quad (2l)$$

The most accurate measurements of refractive index are made by placing the sample in the form of a prism on the table of a spectrometer and measuring the angles α and δ_{min} , the latter for each color desired. When prisms are used in spectroscopes and spectrographs, they are always set as nearly as possible at minimum deviation because otherwise any slight divergence or convergence of the incident light would cause astigmatism

in the image. We have already seen that a divergent pencil incident at any arbitrary angle yields two focal lines T and S . Only at minimum deviation do they merge to form a true point image. To prove this statement we note from Fig. 2J(b) that the widths of the refracted and incident pencils at the prism face are in the ratio

$$\frac{s' \frac{d\phi_2}{d\phi_1}}{s} = \frac{\cos \phi_2}{\cos \phi'_2} \cdot \frac{\cos \phi'_1}{\cos \phi_1} \quad (2m)$$

if we neglect the change in width of the pencil in traversing the prism. That is, the four short segments of wave front shown by broken lines make angles ϕ_1 and ϕ'_1 with the first surface, ϕ_2 and ϕ'_2 , with the second, and we assume the two middle ones to be equal in length. The first focal line T lies at approximately the same distance s from the prism as does the source, because it is the focus for rays in a plane perpendicular to the figure and we are neglecting the thickness of the prism. The ratio of the distances of the two focal lines therefore becomes, from Eqs. 2m and 2i,

$$\frac{s'}{s} = \frac{d\phi_1}{d\phi_2} \cdot \frac{\cos \phi_2}{\cos \phi_1} \cdot \frac{\cos \phi'_1}{\cos \phi'_2} = \left(\frac{d\phi_1}{d\phi_2} \right)^2 \quad (2n)$$

But at minimum deviation $d\phi_1/d\phi_2$ was proved at the end of the last section to be unity. Consequently $s'/s = 1$, or $s = s'$, and the *astigmatism vanishes at minimum deviation*. It also vanishes for parallel light incident at any angle, and therefore it is customary to render the incident beam parallel by means of a *collimating lens* and to focus the emergent parallel light with a *telescope or camera lens*. If the collimator provides strictly parallel light it is permissible to use the prism out of minimum deviation in order, for example, to increase the dispersion.

2.10. Dispersion by a Prism. As was mentioned in Sec. 1.11, the fact that the refractive index depends on color leads to a separation of any incident ray into its component colors when it is refracted through an appreciable angle. The dependence of n' on color cannot be accurately represented by any simple equation, and this question is to be investigated in detail later (Chap. 23). If one knows the value of n' for any one color, a ray of that color may be traced through the prism by means of the relations

$$\left. \begin{aligned} \sin \phi_1 &= n' \sin \phi'_1 \\ \phi'_1 + \phi'_2 &= \alpha \\ n' \sin \phi'_2 &= \sin \phi_2 \end{aligned} \right\} \quad \cdot \quad \cdot \quad \cdot \quad (2o)$$

At a particular point in the spectrum the dispersion of colors is proportional to $d\phi_2/dn'$, and it is important to know how this varies with the

angle of incidence. By differentiating the three equations 2o with respect to n' , regarding ϕ_1 and α as constants, and by combining the results to eliminate $d\phi_1'/dn'$ and $d\phi_2'/dn'$, one obtains

$$\frac{d\phi_2}{dn'} = \frac{\sin \alpha}{\cos \phi_1' \cos \phi_2} \quad (2p)$$

At minimum deviation $\phi_1' = \alpha/2$ and $\phi_2 = \phi_1$. The numerator may be expanded as $2 \sin (\alpha/2) \cos (\alpha/2)$, and the equation for the dispersion then simplifies to

$$\frac{d\phi_2}{dn'} = \frac{2 \sin (\alpha/2)}{\cos \phi_1} \quad \text{AT MINIMUM DEVIATION} \quad (2q)$$

It is difficult to see from Eq. 2p how the dispersion depends on the angle of incidence, so we give in Fig. 2K a curve for a typical case where

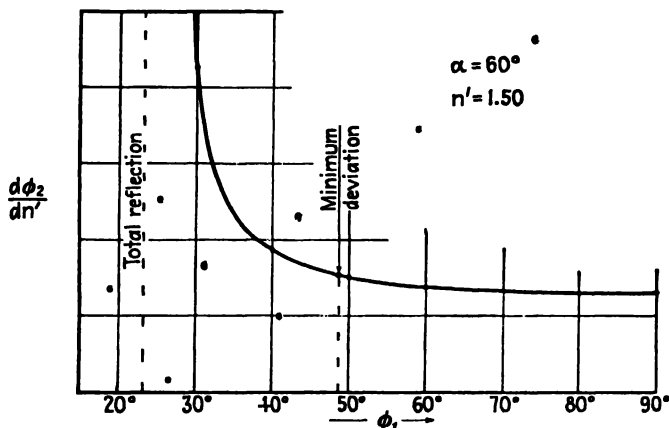


FIG. 2K. Dependence of the dispersion of a prism on the angle of incidence.

$\alpha = 60^\circ$ and $n' = 1.50$. The dispersion is nearly constant in the neighborhood of grazing incidence and increases very little up to the angle for minimum deviation. From there on it grows rapidly, becoming infinite at grazing emergence, as may be seen by putting $\phi_2 = 90^\circ$ in Eq. 2p. Beyond this angle the ray is totally reflected from the second face of the prism, because ϕ_2' exceeds the critical angle. Prisms are not used at angles of incidence much smaller than that corresponding to minimum deviation, even though the dispersion is large. One reason is that a good fraction of the light is lost by internal reflection when the angle at the second face approaches the critical angle.

• For any one angle of incidence on a prism, say that for minimum deviation, the distribution in angle of the other colors will depend upon

the characteristics of the material of the prism. All common optical materials are such that n' increases toward the violet end of the spectrum, and at an increasing rate. Details of this variation will be given in Chap. 23. The result is that the violet is always deviated through a greater angle than the red, and the blue and violet are more widely spread than the orange and red.

As an illustration of the dispersion due to an actual prism, let us compute the angles of deviation of the principal Fraunhofer lines (Sec. 1.11) by a 60° prism of barium flint glass. Assume that the prism is set at minimum deviation for the blue F line, which lies almost in the center

TABLE 2I. DEVIATIONS OF THE FRAUNHOFER LINES BY A 60° PRISM OF FLINT GLASS

Line	n'	ϕ_1	ϕ'_1	ϕ'_2	ϕ_2	δ
C	1.58843	$53^\circ 3'$	$30^\circ 12'$	$29^\circ 48'$	$52^\circ 8'$	$45^\circ 11'$
D	1.59144	$53^\circ 3'$	$30^\circ 9'$	$29^\circ 51'$	$52^\circ 23'$	$45^\circ 26'$
F_2	1.59512	$53^\circ 3'$	$30^\circ 4'$	$29^\circ 56'$	$52^\circ 45'$	$45^\circ 48'$
F	1.59825	$53^\circ 3'$	$30^\circ 0'$	$30^\circ 0'$	$53^\circ 3'$	$46^\circ 6'$
C'	1.63067	$53^\circ 3'$	$29^\circ 53'$	$30^\circ 7'$	$53^\circ 35'$	$46^\circ 38'$
H	1.60870	$53^\circ 3'$	$29^\circ 47'$	$30^\circ 13'$	$54^\circ 3'$	$47^\circ 6'$

of the spectrum. The value of n' for this line is (by Table 2I) 1.59825. Substituting this value in Eq. 2I and using $\alpha = 60^\circ$, we find $\delta_{\min} = 46^\circ 6'$. To find the angle of incidence we then use Eq. 2j with $\phi_1 = \phi_2$, which gives $\phi_1 = \frac{1}{2}(46^\circ 6' + 60^\circ) = 53^\circ 3'$. This angle refers to the incident ray of white light, and is common to all the colors. In finding the angles

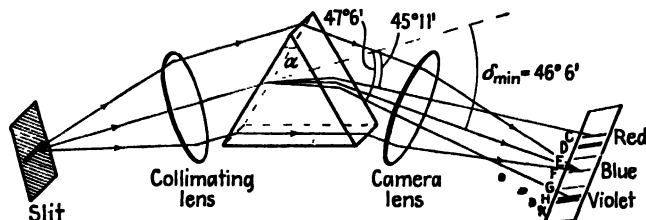


FIG. 2L. Dispersion of the solar spectrum with a simple prism spectrograph.

of emergence ϕ_2 , and hence the total deviations δ , the exact method is to apply the three equations 2o successively for each value of n' . The results of this computation are given in Table 2I. Actually the various colors traverse the prism so nearly at minimum deviation that the application of Eq. 2I gives the δ 's directly, within the accuracy of the figures of Table 2I. Figure 2L represents these results schematically, but the

dispersion had to be greatly exaggerated in order to distinguish the lines. The whole spectrum really should occupy an angle of less than 2° .

The spectrum lines produced by a prism are slightly curved, being concave toward the violet end of the spectrum. This effect is due to the fact that a pencil from one end of the slit does not traverse the prism in a plane exactly perpendicular to the refracting edge. Thus the section of the prism containing such rays has an effective refracting angle larger than α , and the ends of the spectrum lines are more strongly deviated.

Most modern spectrographs utilize prisms of special forms that are

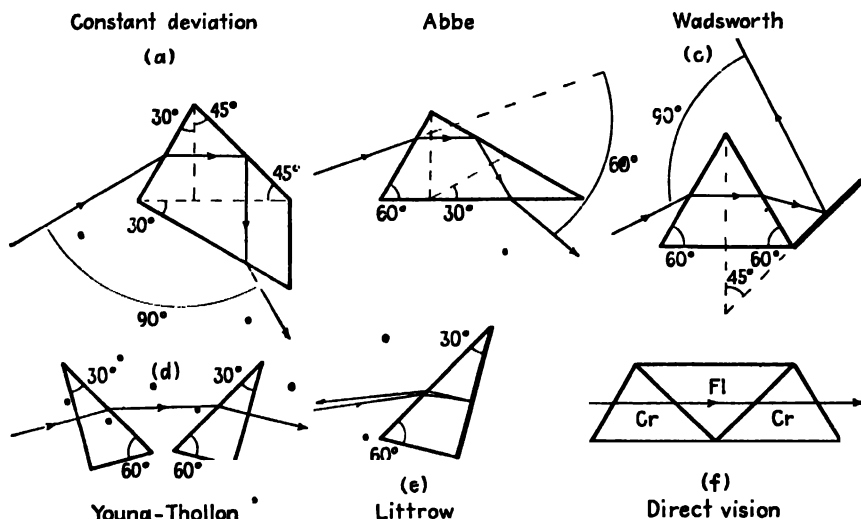


FIG. 2M. Dispersing prisms commonly used in spectrographs and monochromators.

suited to the desired objectives. Figure 2M shows a number of these forms. In (a), (b), and (c) are shown three ways of accomplishing *constant deviation*. It is an interesting problem in geometry to prove that upon rotation of the prism [or prism and mirror in (c)] the deviation of the ray that traverses the effective 60° prism at minimum deviation remains constant and equal to twice the complement of the angle of reflection. This property is very useful in monochromators, which are spectroscopes provided with an exit slit for obtaining nearly monochromatic light. With a fixed angle between collimator and telescope, rotation of the prism brings various wavelengths through the slit, and the images are free from astigmatism since they are always observed at minimum deviation. The Young-Thollon half prisms (d) are frequently used in quartz monochromators for the ultraviolet, one prism being of

left-handed and the other of right-handed quartz (Chap. 27). The Littrow prism (e) is silvered or aluminized on the back surface, and a single lens can be used as both collimator and telescope [principle of autocollimation, as for a grating, in Fig. 17M(c)]. The *direct-vision prism* (f) is made of two crown-glass prisms *Cr* cemented with balsam to a flint-glass prism *Fl*. It produces a spectrum the center of which is undeviated as a result of the fact that the deviations by the crown components are compensated by an opposite deviation by the flint components. Under this condition the dispersions do not cancel, because the dispersive powers of crown and flint glass are different (Sec. 1.11).

2.11. Thin Prism. The equations for the prism become much simpler when the refracting angle α becomes small enough so that its sine, and

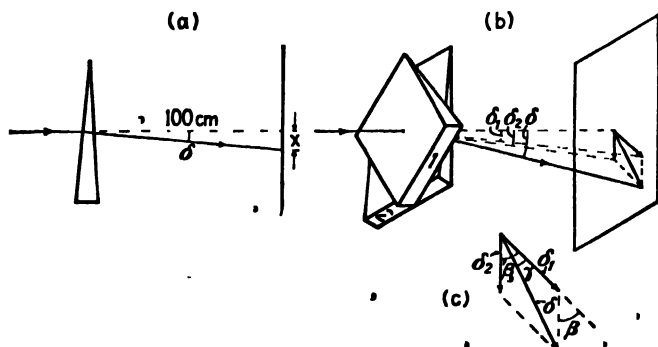


FIG. 2N. Thin prisms. (a) Power x of a single prism; (b) Risley prism of a variable power; (c) vector addition of deviations.

also the sine of the angle of deviation δ , may be set equal to the angles themselves. Even at an angle of 0.1 rad, or 5.7° , the difference between the angle and its sine is less than 0.2 per cent. For prisms having a refracting angle of only a few degrees, we may therefore simplify Eq. 2l by writing

$$n' = \frac{\sin \frac{1}{2}(\delta_{\min} + \alpha)}{\sin \frac{1}{2}\alpha} = \frac{\delta_{\min} + \alpha}{\alpha}$$

and

$$\delta = (n' - 1)\alpha \quad \text{THIN PRISM IN AIR} \quad (2r)$$

The subscript on δ has been dropped because such prisms are always used at or near minimum deviation. It is customary to measure the *power* of a prism by the deflection of the ray in centimeters at a distance of 1 m, in which case the unit of power is called the *prism diopter*. A prism having a power of 1 prism diopter therefore displaces the ray on a screen 1 m away by 1 cm. In Fig. 2N(a) the deflection on the screen, is x cm and is numerically equal to the power of the prism. For small

values of δ it will be seen that the power in prism diopters is essentially the angle of deviation δ measured in units of 0.01 rad, or 0.573° .

For the barium flint glass of Table 2I, $n'_D = 1.59144$, and Eq. 2r shows that the refracting angle of a 1-diopter prism should be

$$\alpha = \frac{0.573}{0.59144} = 0.97^\circ$$

The dispersion of such a prism is negligible, since for the separation of the blue and the red we have

$$\delta_r - \delta_c = (n'_r - n'_c)\alpha = (1.59825 - 1.58848)0.97 = 0.0095^\circ$$

In terms of the dispersive power $1/\nu$ defined in Sec. 1.11, this is just the dispersive power multiplied by the deviation, since by Eq. 2r $\alpha = \delta/(n' - 1)$, so that

$$\delta_r - \delta_c = \frac{n'_r - n'_c}{n'_D - 1} \delta_D = \frac{1}{\nu'} \delta_D \quad (2s)$$

2.12. Combinations of Prisms. In measuring binocular accommodation, ophthalmologists make use of a combination of two thin prisms of equal power which can be rotated in opposite directions in their own plane [Fig. 2N(b)]. Such a device, known as the *Risley or Herschel prism*, is equivalent to a single prism of variable power. When the prisms are parallel the power is twice that of either one, while when they are opposed the power is zero. To find how the power and direction of deviation depend on the angle between the components, we use the fact that the deviations add vectorially. In Fig. 2N(c) it will be seen that the resultant deviation δ will in general be, from the law of cosines,

$$\delta = \sqrt{\delta_1^2 + \delta_2^2 + 2\delta_1\delta_2 \cos \beta} \quad (2t)$$

where β is the angle between the two prisms. To find the angle γ between the resultant deviation and that due to prism 1 alone (or, we may say, between the "equivalent" prism and prism 1) we have the relation

$$\tan \gamma = \frac{\delta_2 \sin \beta}{\delta_1 + \delta_2 \cos \beta} \quad (2u)$$

Since almost always $\delta_1 = \delta_2$, we may call the deviation by either component δ_i , and the equations simplify to

$$\delta = \sqrt{2\delta_i^2(1 + \cos \beta)} = \sqrt{4\delta_i^2 \cos^2 \frac{\beta}{2}} = 2\delta_i \cos \frac{\beta}{2} \quad (2v)$$

and

$$\tan \gamma = \frac{\sin \beta}{1 + \cos \beta} = \tan \frac{\beta}{2}$$

so that

$$\frac{\beta}{2} \quad (2w)$$

There are two important types of prism combination in which the prisms are of different kinds of glass. These are the *direct-vision prism* and the *achromatic prism*. The former was briefly described at the end of Sec. 1.10 and has as its purpose the production of dispersion without deviation. The latter type produces deviation without dispersion. If we can consider the prisms as thin, it is not difficult to derive the neces-

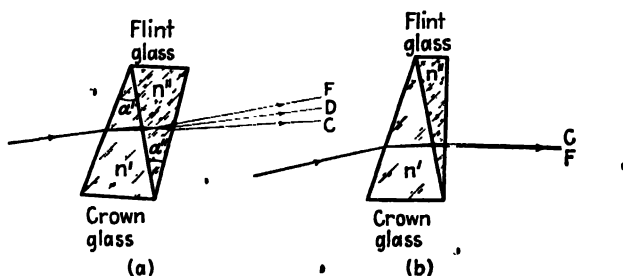


FIG. 20. Prism combinations. (a) Direct-vision prism; (b) achromatic prism.

sary conditions for these cases, and equations for the resultant dispersion or deviation.

For the direct-vision prism [Fig. 20(a)] which consists of a crown-glass prism, index n' and angle α' , opposed by a flint-glass one, index n'' and angle α'' , let us require that the deviation of the D line be zero. Thus $\delta_D = \delta'_D - \delta''_D = 0$. With the help of Eq. 2r, we then find

$$(n'_D - 1)\alpha' - (n''_D - 1)\alpha'' = 0 \quad \frac{\alpha'}{\alpha''} = \frac{n''_D - 1}{n'_D - 1} \quad \text{DIRECT-VISION PRISM} \quad (2x)$$

The residual dispersion is then $\delta_F - \delta_C = (\delta'_F - \delta'_C) - (\delta''_F - \delta''_C)$, which may be written, using Eq. 2s,

$$\delta_F - \delta_C = \frac{1}{v'} \delta'_D - \frac{1}{v''} \delta''_D = \frac{1}{v'} (n' - 1)\alpha' - \frac{1}{v''} (n'' - 1)\alpha''$$

But by Eq. 2x we have $(n'_D - 1)\alpha' = (n''_D - 1)\alpha''$, so that the dispersion becomes

$$\delta_F - \delta_C = (n'_D - 1)\alpha' \left(\frac{1}{v'} - \frac{1}{v''} \right) \quad \text{DIRECT-VISION PRISM} \quad (2y)$$

For the achromatic prism [Fig. 20(b)] we wish the same deviation for all colors. This will be nearly the same if we make it the same for any two, and we therefore take $\delta_r - \delta_c = 0$. Again making use of Eqs. 2r and 2s, equations corresponding to Eqs. 2x and 2y may be derived. These are

$$\frac{\alpha'}{\alpha''} = \frac{n_r'' - n_c''}{n_r' - n_c'}, \quad \delta_{\text{A}} = (n_r' - n_c')\alpha'(\nu' - \nu'') \quad \text{ACHROMATIC PRISM} \quad (2z)$$

These relations for thin prisms are not often used in practice because in the case of the direct-vision prism the dispersion produced by thin prisms is too small and in the case of the achromatic prism there is no need for achromatization unless the angles are fairly large. A prism like that illustrated in Fig. 2M(f) would require a more involved calculation. The thin-prism equations do, however, illustrate the important principle that, by properly adjusting the ratio of the refracting angles of two opposed prisms of different kinds of glass, it is possible to eliminate either the dispersion or the deviation and leave some of the other. Since when $\nu' = \nu''$ both the expressions 2y and 2z are equal to zero, one sees that it is only by virtue of the difference in the dispersive powers of different glasses that these prism combinations can function. Isaac Newton made the error of assuming that the dispersion was exactly proportional to the deviation in different materials (*i.e.*, that they have equal ν 's) and thereby concluded that it was impossible to make an achromatic lens. We shall see in Sec. 9.10 that such lenses are made and that the principle of their design is similar to that of the achromatic prism set forth above.

Problems

1. Prove Eq. 2c for the lateral displacement of a ray which is incident on a plane-parallel plate at an angle ϕ .

2. Carbon disulfide ($n = 1.6255$) is poured into a beaker to a depth of 3 cm. On top of this is poured a layer of water ($n' = 1.3330$) 2 cm deep. (a) How far above or below its true position does a speck of dirt on the bottom of the beaker appear to an observer looking straight down? (b) What is the critical angle for total reflection at the interface between water and carbon disulfide, and from which side of the interface must the light approach?

3. Show that in the refraction of a narrow pencil by a plane surface at any angle the distance from the point of incidence on the surface to the first focal line [T' in Fig. 2G(a)] remains constant and equal to the distance from the surface of the point image formed by paraxial rays at normal incidence.

4. A determination of the index of refraction of lucite is made with a Pulfrich refractometer in which the prism has $n = 1.650$ and a refracting angle $\alpha = 80^\circ$ (see Fig. 2C). The boundary between light and dark is found to occur at an angle of $26^\circ 38'$ with the normal to the second face. Find the refractive index of lucite.

5. A 60° flint-glass prism has refractive indices of 1.75124 and 1.75094 for the two yellow mercury lines. (a) If the prism is used at an angle of incidence $\phi_1 = 70^\circ$ with

a collimator lens of 20 cm focus and a slit 0.15 mm wide, will the images of the yellow lines be separated? (b) If the collimator lens is removed, what will be the distance of the second focal line S from the prism?

6. Show that, for any angle of incidence on a prism,

$$\frac{\sin \frac{1}{2}(\alpha + \delta)}{\sin \frac{1}{2}\alpha} = n' \frac{\cos \frac{1}{2}(\phi'_1 - \phi'_2)}{\cos \frac{1}{2}(\phi_1 - \phi_2)}$$

which reduces to n' at minimum deviation.

7. Two thin prisms are superimposed so that their deviations are at right angles to each other. If the powers are 3 prism diopters and 4 prism diopters, find (a) the resultant deviation in degrees, (b) the power of the resultant prism, and (c) the angle that the resultant prism makes with the weaker of the two prisms.

8. Prove that as long as the angles of incidence and refraction are small enough so that the angles may be substituted for the sines, *i.e.*, for a thin prism not too far from normal incidence, the deviation is independent of the angle of incidence and equal to $(n' - 1)\alpha$.

9. On either side of the angle for minimum deviation there must be two angles of incidence for which the deviations are equal. A 60° prism gives a minimum deviation of $37^\circ 11'$. At an angle of incidence of $63^\circ 27\frac{1}{2}'$ the deviation is 40° . Find the other angle of incidence which gives this same deviation.

10. For the Wadsworth prism-mirror combination [Fig. 2M(c)] prove that the total angle of deviation for a ray traversing the prism at minimum deviation is constant. Find the relation between this total angle δ and the angle β that the plane of the mirror makes with the plane bisecting the refracting angle of the prism.

11. A thin prism of barium flint glass has a refracting angle of 6° and is to be combined with one of borosilicate crown glass to produce a combination that is achromatic for the C and F lines. Find the necessary refracting angle for the crown-glass prism. Find also the resultant deviation of the D line. Use Table 23I for refractive indices.

12. The indices of refraction of a 60° prism for the sodium and lithium lines are 1.5170 and 1.5140, respectively. Suppose the prism to be set at minimum deviation for the sodium line, and the angles of deviation of both the sodium and lithium lines measured. What error is made in the value of the refractive index for the latter line if it is computed as though the measured angle were that for minimum deviation?

13. Determine the angle of minimum deviation for a prism of 45° angle, if the glass has an index of refraction of 1.650.

14. A hollow prism of 60° angle, made with glass plates with parallel sides, is filled with carbon disulfide, index 1.759. Calculate the angle of minimum deviation.

15. Solve Prob. 14 if the hollow prism is filled with water of index 1.333.

CHAPTER 3

THIN LENSES

Most lenses have spherical surfaces, and it might therefore seem appropriate to consider next the refraction by a single spherical surface. Such a surface, in contrast to the plane surface considered in the last chapter, is capable of forming a real image. But this type of image is usually produced not by means of a single spherical surface but by a lens, *i.e.*, by a combination of two such surfaces. Therefore, in order to review certain features of image formation and to establish some important definitions, we give in this chapter the elementary facts about image formation by thin lenses. Most of these will already be familiar to those who have studied elementary physics.

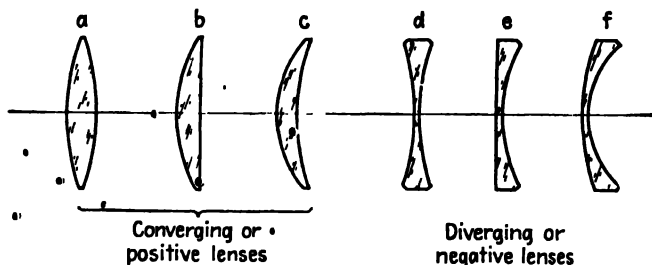


FIG. 3A. Cross sections of common types of thin lenses.

A *thin lens*, as the name implies, is one whose thickness may be neglected so that the vertices of the two surfaces may be regarded as coinciding at the geometrical center of the lens. A more rigorous definition will be given later in Sec. 5.7. Cross sections of several standard forms of thin lenses are shown in Fig. 3A. The three *converging* or *positive lenses*, which are thicker at the center than at the edges, are known as (a) equiconvex, (b) plano-convex, and (c) positive meniscus. The three *diverging* or *negative lenses*, which are thinner at the center, are (d) equiconcave, (e) plano-concave, and (f) negative meniscus. Such lenses are usually made of optical glass as free as possible from inhomogeneities, but occasionally other transparent materials like quartz, fluorite, rock salt, and plastics are used. Although as we shall see the spherical form for the surfaces may not be the ideal one in a particular instance, it gives reasonably good images and is much the easiest to grind and polish.

3.1. Focal Points and Focal Lengths. Diagrams showing the refraction of light by an equiconvex and by an equiconcave lens are given in Fig. 3B. The *axis* in each case is a straight line through the geometrical center of the lens and perpendicular to the two faces at the points of intersection. For spherical lenses this line joins the centers of curvature of the two surfaces. The *primary focal point* F lies on the axis and is defined for a positive lens as the point from which diverging rays are refracted by the lens into a parallel beam. For a negative lens it is the point toward which rays must converge in order to be refracted into a parallel beam. Every thin lens in air has two focal points, one on each side of the lens and equidistant from the center. This may be seen by

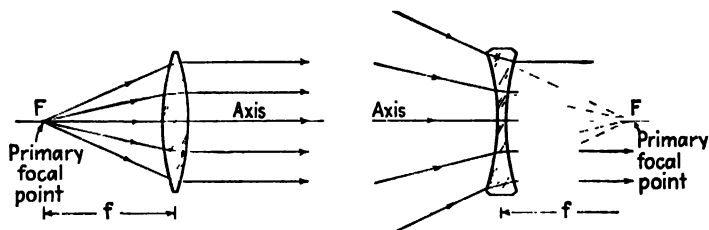


FIG. 3B. Ray diagrams illustrating the primary focal points F and the corresponding primary focal lengths f .

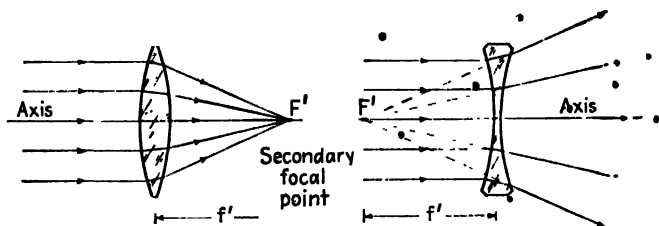


FIG. 3C. Ray diagrams illustrating the secondary focal points F' and the corresponding secondary focal lengths f' .

symmetry in the cases of equiconvex and equiconcave lenses, but it is also true for other forms provided the lenses may be regarded as thin. The *secondary focal point* F' as shown in Fig. 3C may in each case be located by applying the principle of reversibility (Sec. 1.3) to the corresponding diagrams in Fig. 3B. For a positive lens incident parallel rays converge toward F' , while for a negative one they diverge as though they came from F' .

A plane perpendicular to the axis and passing through a focal point is called a *focal plane*. The significance of the focal plane is illustrated for a converging lens in Fig. 3D. Parallel incident rays making an angle θ with the axis are brought to a focus in this plane at a point Q' in line with the *chief ray*. This is defined as the ray which crosses the axis at

some prescribed point—in the present case, at the center of the lens. The distance between the center of a lens and either of its focal points is called its *focal length*. These distances f and f' , usually measured in centimeters or in inches, have a positive sign for converging lenses and a negative sign for diverging lenses. Note in Fig. 3B that the primary focal point F for a converging lens lies to its left, whereas for a diverging

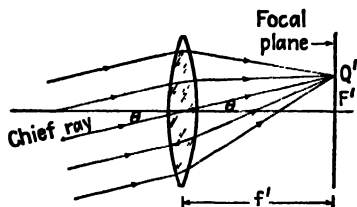


FIG. 3D. Illustrating how parallel incident rays are focused at the focal plane of a lens.

lens it lies to its right. For a lens with the same medium on both sides, we have, by symmetry and by the reversibility of light rays,

$$f = f' \quad \text{LENS IN AIR} \quad (3a)$$

3.2. Image Formation. When an object is placed on either one side or the other of a converging lens and beyond the focal plane, an image is formed on the opposite side. This case is illustrated in Fig. 3E. If the object is moved closer to the focal plane, the image will be formed farther away from the lens and will be larger, *i.e.*, magnified. If the object is moved farther away from the lens, the image will be formed closer to the lens and will be smaller in size.

In Fig. 3E all the rays coming from an object point Q are shown as brought to a focus at Q' and the rays from another point M as brought to a focus at M' . This ideal condition never holds exactly for any actual

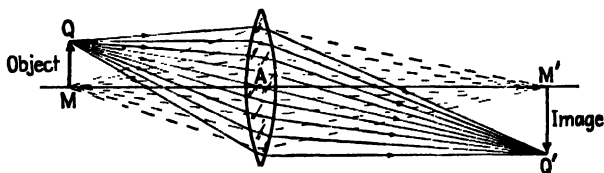


FIG. 3E. Image formation by an ideal thin lens. All rays from an object point Q which pass through the lens are refracted so they pass through the same image point.

spherical lens. Departures from it give rise to defects of the image known as *aberrations*. The elimination of aberrations is the major problem of geometrical optics and will be discussed in some detail later (Chap. 9). If the rays considered are restricted to *paraxial rays*, a good image is formed with monochromatic light, just as in the case of the plane surface discussed in Sec. 2.5. In the case of a lens or lens system, *paraxial rays* are defined as those rays which make very small angles with the axis and lie close to the axis throughout the distance from object to image.

The formulas given in this chapter are to be taken as applying to images formed by paraxial rays.

3.3. Conjugate Points and Planes. The principle of reversibility of light rays has the consequence that if $Q'M'$ in Fig. 3E were an object, an image would be formed at QM . Hence if any object is placed at the position previously occupied by its image it will be imaged at the position previously occupied by the object. The object and image are thus interchangeable, or *conjugate*. Any pair of object and image points such as M and M' in Fig. 3E are called conjugate points, and planes through these points perpendicular to the axis are called conjugate planes.

If one is given the focal length of a thin lens and the position of the object, there are in general two methods of determining the position and size of its conjugate image. One is by graphical construction and the other is by calculation using the lens formula

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \quad (3b)$$

and the corresponding equation for magnification, Eq. 3c below. Here s is the object distance, s' the image distance, and f the focal length, all

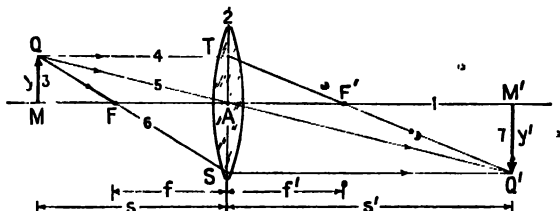


FIG. 3F. Parallel-ray method for the graphical location of an image. f = primary focal length, f' = secondary focal length.

measured to or from the center of the lens. These equations will be derived in Sec. 3.13. Let us consider the graphical methods first.

3.4. Parallel-ray Method. The most familiar graphical method is illustrated in Fig. 3F. Consider the light emitted from the extreme point Q on the object. Of the rays emanating from this point in different directions the one QT , traveling parallel to the axis, will by definition of the focal point be refracted to pass through F' . The ray QA , which goes through the lens center where the faces are parallel, is undeviated and meets the other ray at some point Q' . These two rays are sufficient to locate the tip of the image at Q' , and the rest of the image lies in the conjugate plane through this point. All other rays from Q will also be brought to a focus at Q' . As a check, we note that the ray QF which passes through the primary focal point will by definition of F be refracted parallel to the axis and will cross the others at Q' as shown in the figure.

This method of constructing the image is termed the *parallel-ray method*. The numbers 1, 2, 3, etc., in Fig. 3F indicate the order in which the lines are customarily drawn.

3.5. Oblique-ray Method. The above construction requires the use of an object point which lies off the axis. A somewhat more general graphical method allows us to determine the position of the point that is conjugate to an object point on the axis. This involves finding the place where the ray again crosses the axis, because all other rays from the object point will cross it at that same point. In Fig. 3G let MT represent any ray incident on the lens from the left. It is refracted in the direction TX and crosses the axis at M' . The point X is located as the intersection between the secondary focal plane $F'W$ and the dashed line RR' drawn through the center of the lens parallel to MT .

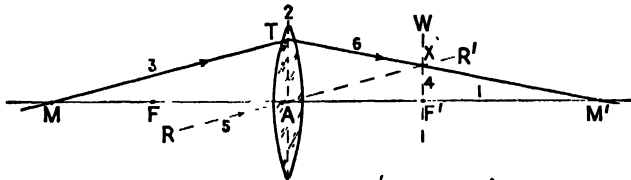


FIG. 3G. Oblique-ray method for the graphical location of an image.

The order in which each step of the construction is made is again indicated by the numbers 1, 2, 3, etc. The principle involved in this method may be understood by reference to Fig. 3D. Parallel rays incident on the lens are brought to a focus at the focal plane, the ray through the lens center being undeviated. Therefore if we actually have rays diverging from M , as in Fig. 3G, we may find the direction of any one of them after it passes through the lens by making it intersect the parallel line RR' through A in the focal plane. This locates X and the position of the image M' . Note that RR' is not an actual ray in this case and is treated as such only as a means of finding the point X .

3.6. Use of the Lens Formula. In illustrating the application of Eq. 3b to find the image position, we take a specific example in which all quantities occurring in the equation have a positive sign. Let an object be located 6 cm in front of a positive lens of focal length 4 cm. Substituting in Eq. 3b, we have

$$\frac{1}{6} + \frac{1}{s'} = \frac{1}{4}$$

Transposing and finding a common denominator, there results

$$\frac{1}{s'} = \frac{1}{4} - \frac{1}{6} = \frac{3}{12} - \frac{2}{12} = \frac{1}{12} \quad \text{or} \quad s' = 12 \text{ cm}$$

The image is formed 12 cm from the lens, and is *real* as it will always be when s' has a positive sign. In this instance it is *inverted*, corresponding to the diagram of Fig. 3E (see also Sec. 3.8 for the analytical method of determining this). These results can be readily checked by either of the two graphical methods presented above.

3.7. Convention of Signs. The following set of sign conventions will be adhered to throughout the following chapters on geometrical optics, and it would be well to have them firmly in mind:

1. All figures are drawn with the light traveling from left to right.
2. All object distances (s) are considered as positive when they are measured to the left of the lens, and negative when they are measured to the right.
3. All image distances (s') are positive when they are measured to the right of the lens, and negative when to the left.
4. Both focal lengths are positive for a converging lens and negative for a diverging lens.
5. Transverse directions are positive when measured upward from the axis and negative when measured downward.

With regard to the signs of radii of curvature, see Sec. 3.10.

3.8. Magnification. In any optical instrument the ratio between the transverse dimension of the final image and the corresponding dimension of the original object defines the *lateral magnification*. A simple formula for the magnification by a single lens may be derived from the geometry of Fig. 3F. By construction it is seen that the right triangles QMA and $Q'M'A$ are similar. Corresponding sides are therefore proportional to each other, so that

$$\frac{M'Q'}{MQ} = \frac{AM}{AM'}$$

where AM' is the image distance s' and AM is the object distance s . Taking upward directions as positive, $y = MQ$ and $y' = -Q'M'$, so we have by direct substitution $y'/y = -s'/s$. The lateral magnification is therefore

$$m = \frac{y'}{y} = -\frac{s'}{s} \quad (3c)$$

When s and s' are both positive, as in Fig. 3F, the negative sign of the magnification signifies an inverted image.

3.9. Virtual Images. The images formed by the converging lenses in Figs. 3E and 3F are real in that they can be made visible on a screen. They are characterized by the fact that rays of light are actually brought

to a focus in the plane of the image. A virtual image (Sec. 2.3) cannot be formed on a screen. The rays from a given point on the object do not actually come together at the corresponding point in the image; instead they must be projected backward to find this point. Virtual images are produced with converging lenses when the object is placed closer to the lens than the focal point, and by diverging lenses when the object is in any position. Examples of these cases are shown in Figs. 3H and 3I.

Figure 3H shows the parallel-ray construction for the case where a positive lens is being used as a magnifier, or reading glass. Rays emanating from Q are refracted by the lens but are not sufficiently deviated to come to a real focus. To the observer's eye at E these rays appear to be com-

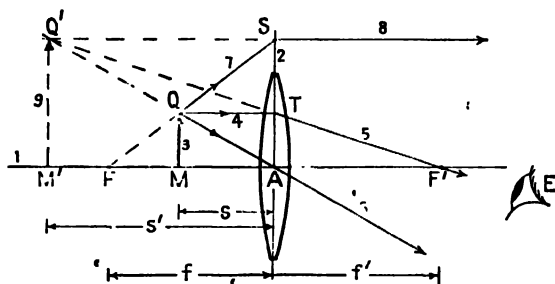


FIG. 3H. Parallel-ray method of construction for the location of the virtual image formed by a positive lens when the object lies inside of the focal point.

ing from a point Q' in back of the lens. This represents a virtual image, because the rays do not actually pass through Q' ; they only appear to come from there. Here the image is right side up and magnified. In the construction of this figure, ray QT parallel to the axis is refracted through F' , while ray QA through the center of the lens is undeviated. These two rays when extended backward intersect at Q' . The third ray QS , traveling outward as though it came from F , actually misses the lens, but if the latter were larger the ray would be refracted parallel to the axis, as shown. When projected backward it also intersects the other projections at Q' .

As an example consider an object placed 6 cm in front of a converging lens of focal length 10 cm, and let it be required to find the image. Direct substitution in Eq. 3b gives

$$\frac{1}{6} + \frac{1}{s'} = \frac{1}{10} \qquad \frac{1}{10} \qquad \frac{3}{30} \qquad \frac{5}{30} \qquad \frac{2}{30}$$

from which $s' = -15\text{cm}$. The minus sign indicates a virtual image to the left of the lens. By substitution in Eq. 3c the magnification is

$$m = -\frac{s'}{s} = -\frac{-15}{6} = +2.5$$

The positive sign means that the image is erect.

In the case of the negative lens shown in Fig. 3I the image is virtual for all positions of the object, is always smaller than the object, and lies closer to the lens than does the object. As is seen from the diagram, rays diverging from the object point Q are made more divergent by the lens. To the observer's eye at E these rays appear to be coming from the point Q' in back of but close to the lens. In applying the lens

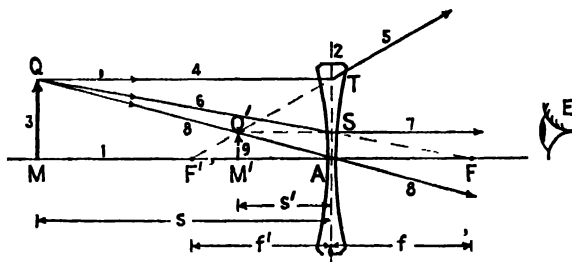


FIG. 3I. Parallel-ray method of construction for the location of the virtual image formed by a negative lens. (NOTE: The order of the primary and secondary focal points is just the reverse of that for a positive lens.)

formula to a diverging lens it must be remembered that the focal length f is negative.

To illustrate, consider the following example:

Example: An object is placed 12 cm in front of a diverging lens of focal length 6 cm. Find the image.

Solution: Again making use of Eq. 3b, we obtain

$$\frac{1}{12} - \frac{1}{-6} = \frac{1}{s'} \quad \frac{1}{s'} = -\frac{3}{12}$$

from which $s' = -\frac{12}{3} = -4\text{ cm}$. For the image size, Eq. 3c gives

$$m = -\frac{s'}{s} = -\frac{-4}{12} = +\frac{1}{3}$$

The image is therefore to the left of the lens, virtual, erect, and one-third the size of the object.

3.10. Lens Makers' Formula. If a lens is to be ground to some specified focal length, the refractive index of the glass must be known. It is customary for manufacturers of optical glass to specify the refractive index for the yellow sodium D line. Supposing the index to be known, the radii of curvature must be so chosen as to satisfy the equation

$$\frac{1}{f} = (n - 1) \left(\frac{1}{r_1} - \frac{1}{r_2} \right) \quad (3d)$$

As the rays travel from left to right through a lens, *all convex surfaces encountered are taken as having a positive radius, and all concave surfaces encountered, a negative radius.* For an equiconvex lens like the one in Fig. 3A(a), r_1 for the first surface is positive and r_2 for the second surface negative. Substituting the value of $1/f$ from Eq. 3b, we may write

$$\frac{1}{s} + \frac{1}{s'} = (n - 1) \left(\frac{1}{r_1} - \frac{1}{r_2} \right) \quad (3e)$$

Example: A plano-convex lens having a focal length of +25 cm [Fig. 3A(b)] is to be made of glass of refractive index $n = 1.520$. Calculate the radius of curvature of the grinding and polishing tools that must be used to make this lens.

Solution: Since a plano-convex lens has one flat surface, the radius for that surface is infinite, and r_1 in Eq. 3d is replaced by ∞ . The radius r_2 of the second surface is the unknown. Substitution of the known quantities in Eq. 3d gives

$$\frac{1}{25} = (1.520 - 1) \left(\frac{1}{\infty} - \frac{1}{r_2} \right)$$

Transposing and solving for r_2 ,

$$\begin{aligned} \frac{1}{25} &= 0.520 \left(0 - \frac{1}{r_2} \right) = - \frac{0.520}{r_2} \\ r_2 &= - (25 \times 0.520) = -13.0 \text{ cm} \end{aligned}$$

If this lens is turned around, as in the figure, we will have $r_1 = +13.0$ cm and $r_2 = \infty$.

3.11. Lens Power in Diopters. It is customary in optometry and ophthalmology to specify lenses not by their focal length but by their *power* in diopters. This is a quantity analogous to the power of a prism

(Sec. 2.11). The power of a lens in diopters is given by *the reciprocal of the focal length in meters*. Thus

$$P = \frac{1}{f} \quad \text{Diopters} = \frac{1}{\text{focal length in meters}} \quad (3f)$$

For example, a lens with a focal length of +50 cm has a power of $1/0.50 = +2$ diopters, ($P = +2$ D), whereas one of -20 cm focal length has a power of $1/-0.20 = -5$ diopters, ($P = -5$ D), etc. Converging lenses have a positive power while diverging ones have a negative power. By Eq. 3f the power of a lens has the dimensions of 1/meters, or meters⁻¹.

Spectacle lenses are made to the nearest quarter of a diopter, thereby reducing the number of grinding and polishing tools required in the optical shops. Furthermore, the sides next to the eyes are always hollow ground to permit free movement of the eyelashes and yet to keep the lens as close to and as normal to the axis of the eye as possible.

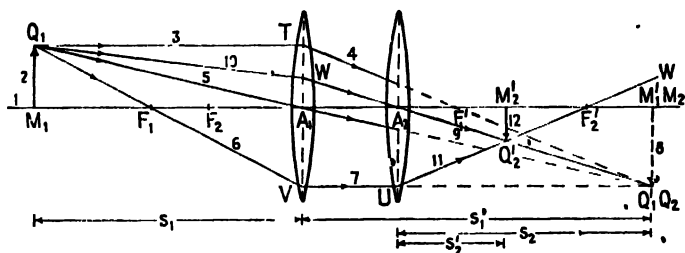


FIG. 3J. Parallel-ray method of locating the image formed by two thin lenses.

NOTE: It is important to insert a *plus or minus sign* in front of the number specifying lens power, thus, $P = +3$ D, $P = -4.5$ D, etc.

3.12. Thin-lens Combinations. The graphical and also the analytical methods of finding the image which have been presented above are readily extended to optical systems involving two or more thin lenses. Consider, for example, two converging lenses spaced some distance apart, as shown in Fig. 3J. When the focal lengths of the lenses and their separation are known, the final image of an object such as Q_1M_1 is located in the same manner as was done in Sec. 2.7 for the image seen through a plane-parallel plate.

As a first step the right-hand lens is disregarded, and the image produced by the first lens alone is determined. In the diagram the parallel-ray method as applied to the object point Q_1 locates a real inverted image at Q'_1 . This image then becomes the object for the second lens, and hence is also labeled Q_2 . Since all possible rays from Q_1 as they leave the first lens must be converging toward Q'_1 , the ray $Q'_1A_2WQ_1$ can

be drawn in reverse as the one ray that passes undeviated through the second lens. Since ray 7 between the lenses is parallel to the axis, it will upon refraction by the second lens pass through F'_2 . The intersection of ray 9 and ray 11 locates the final image point Q'_2 . M_1 and M'_1 are conjugate points for the first lens, and M_2 and M'_2 are conjugate for the second lens, while M_1 and M'_2 are conjugate for the combination of two lenses.

The oblique-ray method as applied to the same two lenses is shown in Fig. 3K. A single ray is traced through the combination and yields the same set of conjugate points along the axis.

By way of comparison and as a check on the graphical solutions, the lens formula, Eq. 3b, will be applied to the following problem: Two converging lenses having focal lengths $f_1 = +3$ cm and $f_2 = +4$ cm are

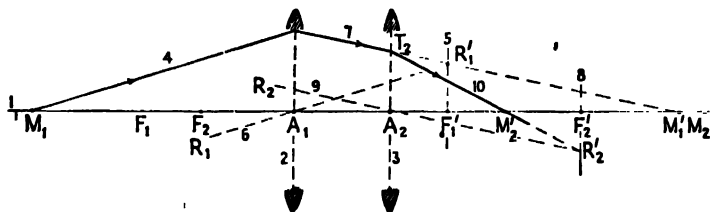


FIG. 3K. The oblique-ray method of locating the position of the image with two thin lenses.

placed 2 cm apart. An object is located 4 cm in front of the first lens. Find the final image.

To solve this problem, we begin by applying the formula to the first lens alone, obtaining

$$\frac{1}{4} + \frac{1}{s'_1} = \frac{1}{3}, \quad \frac{1}{s'_1} = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

from which $s'_1 = 12$ cm. Since this image Q'_1 is 12 cm to the right of the first lens, it is only 10 cm from the second lens. The object distance for the second lens is therefore $s_2 = -10$ cm. The negative sign is essential, and results from the fact that the object distance is measured to the right of the second lens. We say that the image which the first lens would give in the absence of the second lens constitutes a *virtual object* for the second lens. Applying the lens formula to the second lens, we obtain

$$\frac{1}{10} + \frac{1}{s'_2} = \frac{1}{4}, \quad \frac{1}{s'_2} = \frac{1}{4} - \frac{1}{10} = \frac{7}{20}$$

giving $s'_2 = \frac{20}{7} = 2.86$ cm.

3.13. Derivation of the Lens Formula. A derivation of Eq. 3b, which is the lens formula in the *Gaussian* form*, is readily obtained from the geometry of Fig. 3L. The necessary features are repeated in Fig. 3L, which shows only two rays leading from the object of height y to the image of height y' . Let s and s' as usual represent the object and image distances from the lens center, and x and x' their respective distances measured from the focal points F and F' .

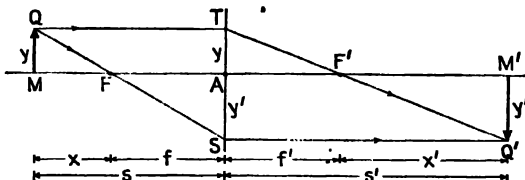


FIG. 3L. Geometrical diagram from which the lens formulas may be derived.

From the similar triangles $Q'TS$ and $F'TA$ the proportionality between corresponding sides gives

$$\frac{y - y'}{s'} = \frac{y}{f'}$$

Note that $y - y'$ is written instead of $y + y'$, because y' by the convention of signs is a negative quantity. From the similar triangles $Q'TS$ and FAS ,

$$\frac{y - y'}{s} = -\frac{y'}{f}$$

The sum of these two equations is

$$\frac{y - y'}{s} + \frac{y - y'}{s'} = \frac{y}{f'} - \frac{y'}{f}$$

Since in air $f = f'$, the two terms on the right may be combined and $y - y'$ canceled out, yielding the desired result,

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \quad \text{GAUSSIAN LENS FORMULA} \quad (3b)$$

We shall derive the lens makers' formula later in Sec. 4.9.

* Karl Friedrich Gauss (1777-1855). German astronomer and physicist, chiefly known for his contributions in the mathematical theory of magnetism. Coming from a poor family, he received support for his education because of his obvious mathematical ability. In 1841 he published the first general treatment of the first-order theory of lenses in his now famous papers, "Dioptrische Untersuchungen."

Another form of the lens formula, the so-called *Newtonian form*, is obtained in an analogous way from two other sets of similar triangles, MQF and FAS on the one hand, and TAF' and $F'M'Q'$ on the other. We find

$$\frac{y}{x} = -\frac{y'}{f} \quad \text{and} \quad -\frac{y'}{x'} = \frac{y}{f} \quad (3g)$$

Multiplication of one equation by the other gives

$$xx' = f^2 \quad \text{NEWTONIAN LENS FORMULA} \quad (3h)$$

In the Gaussian formula the object and image distances are measured from the lens, while in the Newtonian formula they are measured from the focal points. Object distances (s or x) are positive if the object lies to the left of its reference point (A or F , respectively), while image distances (s' or x') are positive if the image lies to the right of its reference point (A or F' , respectively).

The lateral magnification as given by Eq. 3c corresponds to the Gaussian form. When distances are measured from focal points, one must use the Newtonian form, which may be obtained at once from Eqs. 3g as

$$m = \frac{y'}{y} = -\frac{f}{x} = -\frac{x'}{f} \quad \text{LATERAL MAGNIFICATION (NEWTONIAN)} \quad (3i)$$

as compared to

$$m = \frac{y'}{y} = -\frac{s'}{s} \quad \text{LATERAL MAGNIFICATION (GAUSSIAN)} \quad (3c)$$

3.14. Object Space and Image Space. For every position of the object there is a corresponding position for the image. Since the image may be either real or virtual and may lie on either side of the lens, the *image space* extends from infinity in one direction to infinity in the other. But object and image points are conjugate, so the same argument holds for the *object space*. In view of their complete overlapping, one might wonder how it is that the distinction between object and image space is made. This is done by defining everything that pertains to the rays before they have passed through the refracting system as belonging to the object space, and everything that pertains to them afterward as belonging to the image space. Referring to Fig. 3J, the object Q_1 and the rays Q_1T , Q_1A_1 , and Q_1V are all in the object space for the first lens. Once these rays leave that lens, they are in the image space of the first lens, as is also the image Q'_1 . This space is also the object space for the second lens. Once the rays leave the second lens, they and the image point Q'_2 are in the image space of the second lens.

3.15. Thin Lenses in Contact. When two thin lenses are placed in contact as shown in Fig. 3M, the combination will act as a single lens with two focal points symmetrically located at F and F' on opposite sides. Parallel incoming rays are shown refracted by the first lens toward its secondary focal point F'_1 . Further refraction by the second lens brings the rays together at F' . This latter is defined as the secondary focal point of the combination, and its distance from the center is defined as the combination's focal length f .

If we now apply the simple lens formula, Eq. 3b, to the rays as they enter and leave the second lens L_2 , we note that for the second lens alone

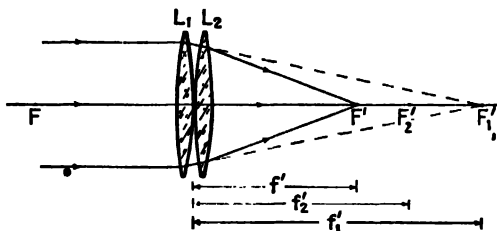


FIG. 3M. The power of a combination of thin lenses in contact is equal to the sum of the powers of the individual lenses.

f_1 is the object distance (taken with a negative sign), f is the image distance, and f_2 is the focal length. Applying Eq. 3b, these substitutions for s , s' , and f respectively give

$$-\frac{1}{f_1} + \frac{1}{f} = \frac{1}{f_2}$$

Transposing,

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} \quad (3j)$$

In other words, the reciprocal of the focal length of a thin-lens combination is equal to the sum of the reciprocals of the focal lengths of the individual lenses.

Since by Eq. 3f we can write $P_1 = 1/f_1$, $P_2 = 1/f_2$, and $P = 1/f$, we obtain for the power of the combination,

$$P = P_1 + P_2 \quad (3k)$$

In general *when thin lenses are placed in contact the power of the combination is given by the sum of the powers of the individual lenses.*

Problems

1. An object 5 cm high is placed 18 cm in front of a thin positive lens of focal length 12 cm. Find (a) the image distance, (b) the magnification, and (c) the nature of the image. Solve graphically and by calculation.

2. An object, located 30 cm in front of a thin lens, has its image formed on the opposite side 60 cm from the lens. Calculate (a) the focal length of the lens, and (b) the lens power.

3. An object 12 cm high is located 60 cm in front of a negative lens whose focal length $f = -20$ cm. Calculate (a) the power of the lens, (b) the image distance, and (c) the lateral magnification. Graphically locate the image by (d) the parallel-ray method, and (e) the oblique-ray method.

4. The radii of a thin lens are $r_1 = +10$ cm, and $r_2 = -25$ cm. The lens is made of glass of index 1.600. Calculate (a) the focal length and (b) the power of the lens.

5. A plano-convex lens is to be made of flint glass of index $n = 1.65$. Calculate the radius of curvature necessary to give the lens a power of +4 D.

6. An equiconcave lens is to be made of crown glass of index $n = 1.52$. Calculate the radii of curvature if it is to have a power of -3 D.

7. A converging lens is used to focus the image of a candle flame on a distant screen. A second lens with radii $r_1 = +15$ cm and $r_2 = -30$ cm and index 1.50 is placed in the converging beam 50 cm from the screen. Calculate (a) the power of the second lens, (b) the position of the final image.

8. Two lenses with focal lengths $f_1 = +20$ cm and $f_2 = +30$ cm are located 10 cm apart. If an object 5 cm high is located 30 cm in front of the first lens find (a) the position, and (b) the size of the final image.

9. Two lenses having focal lengths $f_1 = +20$ cm and $f_2 = -30$ cm are placed 20 cm apart. If an object 5 cm high is located 50 cm in front of the first lens, find (a) the position, and (b) the size of the final image.

10. A double-convex lens is to be made of glass having an index of 1.65. If one surface is to have twice the radius of the other and the focal length is to be 20 cm, find the radii.

11. An object is located 1 m from a white screen. What focal length lens will be required to form a real, inverted image on the screen with a magnification of 4?

12. A lantern slide 3 in. high is located 20 ft from a projection screen. What focal length lens will be required to project an image 3 ft high?

13. Two thin lenses having the following radii of curvature and index are placed in contact: $r'_1 = +20$ cm, $r'_2 = -30$ cm, $r''_1 = -40$ cm, $r''_2 = +60$ cm, $n' = 1.50$, $n'' = 1.60$. Find their combined (a) focal length, and (b) power.

14. Three thin lenses have the following powers: -1 D, +2 D, and +4 D. What are all the possible powers obtainable with these lenses using one, two, or three in contact?

CHAPTER 4

SPHERICAL SURFACES

There are close analogies between the effects produced by a thin lens and those produced by a single spherical refracting surface. Such a surface can focus a divergent pencil to form a real image. It therefore has associated with it focal points and focal lengths just as does the thin lens. Although the real or virtual images produced by a single surface also have magnification or reduction, they are seldom made use of directly in optical instruments. Nevertheless, they occur as intermediate stages in the functioning of any lens system involving two or

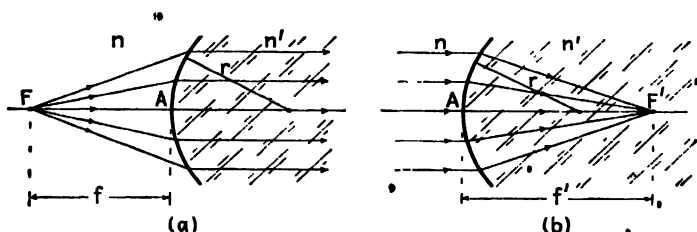


FIG. 4A. Refraction at a single spherical surface showing (a) the primary focal point F and primary focal length f , and (b) the secondary focal point F' and secondary focal length f' .

more spherical surfaces, and it is therefore important to have an understanding of their formation before attempting to analyze the more difficult cases. In this chapter we shall discuss first the behavior of paraxial rays upon refraction at one spherical surface, and then at two in succession. The latter case forms the basis for the treatment of thick lenses to be given in Chap. 5. Reflection at spherical surfaces will then be considered in Chap. 6.

4.1. Foci of a Single Surface. Figure 4A shows a spherical surface of radius r separating two media of indices n and n' . In the first diagram rays diverge from a point source F in the first medium and are refracted into a parallel beam. In the second a parallel beam incident on the same surface is shown as brought to a point focus F' . The similarity between these diagrams and the corresponding ones for a thin converging lens in Figs. 3B and 3C will be obvious. As before, F and F' are called the primary and secondary focal points, and their distances f and f' measured from the vertex A are the primary and secondary focal lengths.

In contrast to the case of the thin lens in air, the two focal lengths are not equal; the ratio of f' to f is found to be n'/n . Incidentally, the same would be true for a lens if it had different media on the two sides.

The significance of the focal plane associated with a single surface is illustrated in Fig. 4B. This figure is closely analogous to Fig. 3D for a thin lens, but there is one important difference. The ray which is undeviated must be the one that goes through the center of curvature

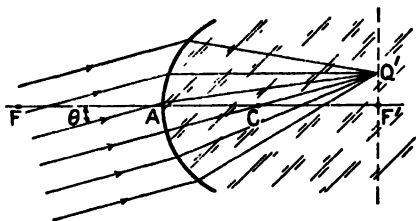


FIG. 4B. Parallel rays are brought to a focus at Q' in the focal plane through F' .

C , because this one meets the surface perpendicularly. Hence the chief ray, which in a thin lens crosses the axis at the center of the lens, here crosses it not at the vertex but at the center of curvature.

4.2. Conjugate Points and Planes.

A diagram illustrating the image formation by a single refracting surface is given in Fig. 4C. It has been drawn for the case in which the first medium is air with an index $n = 1$ and the second medium is glass with an index $n' = 1.60$. The focal lengths f and f' therefore have the ratio 1:1.6.

An equation similar to the lens makers' formula, Eq. 3d, will be derived for a single surface in Sec. 4.8. Such a formula gives the image distance

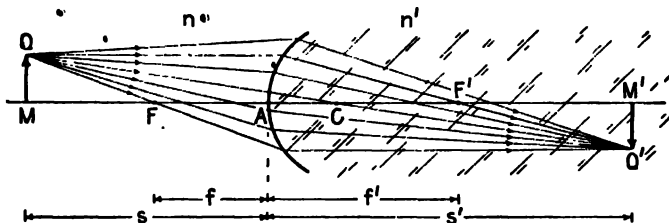


FIG. 4C. All rays leaving the object point Q , and passing through the refracting surface, are brought to a focus at the image point Q' .

s' in terms of the object distance s , the refractive indices n and n' , and the radius of curvature r , as follows:

$$\frac{n}{s} + \frac{n'}{s'} = \frac{n' - n}{r} \quad (4a)$$

Example: The left-hand end of a solid glass rod of index 1.50 is ground and polished to a hemispherical surface of radius 1 cm. A small object is placed in air on the axis 4 cm to the left of the vertex. Find the position of the image. Assume $n = 1.00$ for air.

Solution: Substituting the given quantities in Eq. 4a, one obtains

$$\frac{1}{4} + \frac{1.50}{s'} = \frac{1.50 - 1.00}{1}, \quad \frac{1.50}{s'} = \frac{0.5}{1} - \frac{1}{4}$$

from which $s' = 6$ cm. Applying the same sign conventions as were used for lenses (Sec. 3.7), one concludes that a real image is formed in the glass rod 6 cm to the right of the vertex.

As the object point M is brought closer to the primary focal point, Eq. 4a shows that the image distance AM' becomes steadily greater and that in the limit when the object reaches F' the refracted rays are parallel and the image is formed at infinity. Then we have $s' = \infty$ and Eq. 4a becomes

$$\frac{n}{s} + \frac{n'}{\infty} = \frac{n' - n}{r}$$

Since this particular object distance is called the primary focal length f , we may then write

$$\frac{n}{f} = \frac{n' - n}{r} \quad (4b)$$

Similarly, if the object distance is made larger and eventually approaches infinity, the image distance diminishes and becomes equal to f' in the limit $s = \infty$. Then

$$\frac{n}{\infty} + \frac{n'}{s'} = \frac{n' - n}{r}$$

or, since this value of s' represents the secondary focal length f' ,

$$\frac{n'}{f'} = \frac{n' - n}{r} \quad (4c)$$

Equating the left-hand members of Eqs. 4b and 4c, we get

$$\frac{n}{f} = \frac{n'}{f'} \quad \text{or} \quad \frac{n}{n'} = \frac{f}{f'} \quad (4d)$$

When $(n' - n)/r$ in Eq. 4a is replaced by n/f or by n'/f' according to Eqs. 4b or 4c, there results

$$\frac{n}{s} + \frac{n'}{s'} = \frac{n}{f} \quad \text{or} \quad \frac{n}{s} + \frac{n'}{s'} = \frac{n'}{f'} \quad (4e)$$

Either of these equations gives the conjugate distances for a single spherical surface and is analogous to the Gaussian formula, Eq. 3a, for a thin lens.

4.3. Graphical Constructions. It would be well to remind the reader at this point that, although the above formulas hold for all possible object and image distances, they apply only to images formed by paraxial rays. For such rays the refraction occurs at or very near the vertex of the spherical surface, so that the correct geometrical relations are obtained in graphical solutions by drawing all rays as though they were refracted at the plane through the vertex A and normal to the axis.

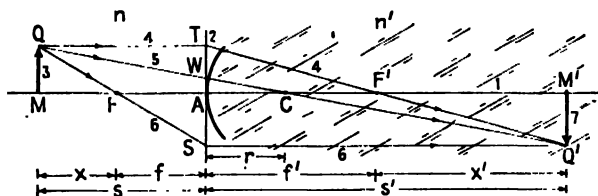


FIG. 4D. Illustrating the parallel-ray method for graphically locating the image formed by a single spherical surface.

The parallel-ray method of construction is illustrated in Figs. 4D and 4E for convex and concave surfaces respectively. These constructions are made in the same manner as those for thin lenses that were illustrated in Figs. 3F and 3I, with the single exception that the undeviated ray 5 is drawn toward the center of curvature C rather than through the central point A . In both these figures the medium to the right of the surface has the greater index; i.e., we take $n' > n$. If in Fig. 4D the medium on the left were to have the greater index, so that $n' < n$, the surface

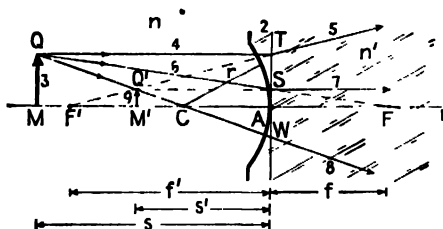


FIG. 4E. Illustrating the parallel-ray method applied to a concave spherical surface having diverging properties.

would have a diverging effect and each of the focal points F and F' would lie on the opposite side of the vertex from that shown, just as they do in Fig. 4E. Similarly, if we had $n' < n$ in Fig. 4E, the surface would have a converging effect and the focal points would lie as they do in Fig. 4D. Similar modifications are obtained in the case of thin lenses if the index of the lens is made smaller than that of the surroundings, as in the case of an "air lens" (enclosed within two watch glasses cemented together) when immersed in water.

The oblique-ray method as applied to a single refracting surface is shown in Fig. 4*F*. Again the construction is very similar to that in the thin-lens case shown in Fig. 3*G*, and identical symbols have been used here. The only difference is that the "hypothetical" ray *R**X* is drawn through the center of curvature *C*, whereas in Fig. 3*G* the corresponding line was drawn through the lens center. Just as before, *M* and *M'* are conjugate points and represent a point object and its image.

In order to perform a graphical solution of any problem, the focal lengths *f* and *f'* either must be given or must first be calculated from the known refractive indices and radius of curvature. For this purpose Eqs. 4*b* and 4*c* are to be used. The student is urged to carry out graphical constructions for various conditions to see the effect of changing the object distance and the relative magnitudes of *n* and *n'*. In particular,

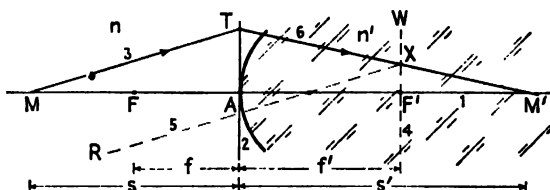


FIG. 4*F*. Illustrating the oblique-ray method for graphically locating image rays and points.

the formation of a virtual image, in analogy to Fig. 3*H*, should be investigated.

4.4. Magnification. To determine the relative size of the image formed by a single surface, reference is made to the geometry of Fig. 4*D*. Here the undeviated ray 5 forms two right triangles *QMC* and *Q'M'C*. Proportionality of corresponding sides requires that

$$\frac{M'Q'}{MQ} = \frac{CM'}{CM} \quad \text{or} \quad -\frac{y'}{y} = \frac{s' - r}{s + r}$$

Since by definition y'/y is the lateral magnification *m*, we obtain

$$m = \frac{y'}{y} = -\frac{s' - r}{s + r} \quad (4f)$$

If *m* is positive the image is virtual and erect, while if it is negative the image is real and inverted.

Another very useful formula for the magnification is derived in Sec. 8.5. By Eq. 8*n*,

$$m = -\frac{ns'}{n's} \quad (4g)$$

4.5. Newtonian Formulas. In the Newtonian form of the lens formula (Eq. 3*h*) the object and image distances x and x' were measured from the focal points rather than from the lens itself. A similar equation may be derived for a single refracting surface. By inspection of Fig. 4*D* one observes that the triangles QMF and SAF form one set of similar triangles, while TAF' and $Q'M'F'$ form another. We may therefore write, for the two sets,

$$\frac{y}{x} = -\frac{y'}{f} \quad \text{and} \quad \frac{y}{f} = -\frac{y'}{x'}$$

Transposing and solving each for the lateral magnification, we find

$$m = \frac{y'}{y} = -\frac{f}{x} = -\frac{x'}{f'}$$

From the last equality we obtain the Newtonian relation analogous to the lens formula, which here becomes

$$xx' = ff' \quad (4h)$$

4.6. Reduced Vergence. In the Gaussian formulas for the lens and for the single surface, Eqs. 3*b*, 3*d*, 4*a*, and 4*e*, the distances s , s' , r , r' , f , and f' appear in the denominators. The reciprocals of these magnitudes

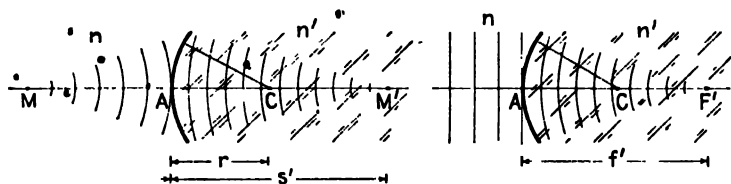


FIG. 4*G*. Illustrating the refraction of light waves by a single spherical surface.

(namely, $1/s$, $1/s'$, $1/r$, $1/r'$, $1/f$, and $1/f'$) actually represent *curvatures*, of which s , s' , r , r' , f , and f' are the radii. As regards the object and image distances s and s' , the curvatures involved are those of *wave fronts* (Sec. 1.8). Reference to Fig. 4*G* will show that the wave fronts in the object space have a radius s and a curvature $1/s$ just as they arrive at the refracting surface. The latter has a radius r and curvature $1/r$. Immediately after refraction the radius of the wave front is s' and its curvature $1/s'$, the center of curvature being at the image point M' . In the second diagram the object point is at infinity and the incident wave fronts have zero curvature ($1/\infty$) as they strike the surface. Their curvature is $1/f$ immediately after refraction. The Gaussian formulas may therefore be considered as involving the addition and subtraction of quantities proportional to the curvatures of convergent and divergent

surfaces. The formulas become simpler and more convenient when these curvatures, rather than radii, are used.

We may now introduce the following quantities:

$$\begin{aligned} V &= \frac{n}{s}, & V' &= \frac{n'}{s'}, & K &= \frac{1}{r} \\ P &= \frac{n}{f}, & \text{and} & & P' &= \frac{n'}{f'} \end{aligned} \quad (4i)$$

The first two of these, V and V' , are called *reduced vergences* because they are direct measures of the convergence (or divergence) of the object and image wave fronts, respectively. For a divergent wave front in the object space, s is positive and so is its vergence V . For a convergent object wave front, on the other hand, s is negative, as is also the vergence. For a converging image wave front V' is positive and for a diverging one V' is negative. Note that in each case the refractive index involved is that of the medium in which the wave front is located.

The third quantity K is the actual curvature of the refracting surface (reciprocal of its radius), while the third and fourth quantities are, according to Eq. 4d, equal, and define its refracting power P . When all distances are measured in meters, the reduced vergences V and V' , the curvature K , and the power P are in diopters (see Sec. 3.11). Remembering that the diopter is a unit of power, we see that V represents the power of the object wave front that just touches the refracting surface (of curvature K and power P), while V' represents the power of the corresponding image wave front which is tangent to that refracting surface. In these new terms, Eq. 4a becomes

$$V + V' = (n' - n)K \quad (4j)$$

Eq. 4b becomes

$$P = \frac{n' - n}{r} \quad \text{or} \quad P = (n' - n)K \quad (4k)$$

and Eq. 4c becomes

$$V + V' = P \quad (4l)$$

Example: One end of a glass rod of refractive index 1.50 is ground and polished with a convex spherical surface of radius 10 cm. An object is placed in air on the axis 40 cm to the left of the vertex. Find (a) the curvature of the surface, (b) the power of the surface, and (c) the position of the image.

Solution: (a) From Eq. 4i, expressing distances in meters, we have

$$K = \frac{1}{r} = \frac{1}{0.10} = +10 \text{ D}$$

(b) Using Eq. 4k,

$$P = (1.50 - 1.00) \times 10 = +5 \text{ D}$$

(c) For the powers of the wave fronts, we have, from Eq. 4i,

$$V = \frac{1.00}{0.40} = +2.5 \text{ D}$$

and from Eq. 4l,

$$\begin{aligned} 2.5 + V' &= +5 \\ V' &= +2.5 \text{ D} \end{aligned}$$

To find the image distance, we have $V' = n'/s'$, so that

$$s' = \frac{n'}{V'} = \frac{1.50}{2.5} = 0.60$$

or

$$s' = 60 \text{ cm}$$

This answer should be verified by the student, using one of the graphical constructions drawn to a convenient scale.

4.7. Two Spherical Surfaces. When two spherical surfaces are involved in the formation of an image, the problem may be solved by the use of the preceding formulas, applying them first to one surface alone and then to the other. The procedure is much the same as that discussed in Sec. 3.12 for two thin lenses; the image for the first refracting surface becomes the virtual object for the second.

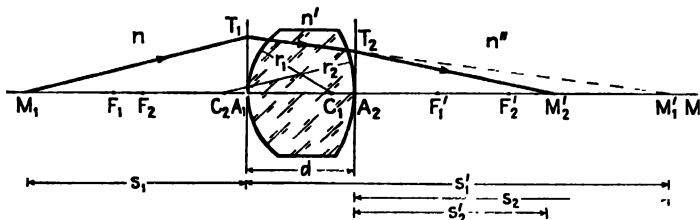


FIG. 4H. Refraction at two spherical surfaces (thick lens).

To illustrate, consider the geometry of Fig. 4H, where two surfaces of radii r_1 and r_2 separate three media of refractive indices n , n' , and n'' respectively. The oblique ray M_1T_1 from the object point M_1 is refracted toward the conjugate image point M_1' . Upon reaching T_2 the ray is further refracted from its direction T_2M_1' to the new direction T_2M_2' . In other words, M_1' is also the object point M_2 for the second surface, and M_2' is its conjugate image point.

Example: An equiconvex lens 2 cm thick and having radii of curvature 2 cm is mounted in the end of a water tank. An object in air is placed

on the axis of the lens 5 cm from its first vertex. Find the position of the final image. Assume refractive indices of 1.00, 1.50, and 1.33 for air, glass, and water, respectively.

Solution: The relative dimensions in this problem are those shown in Fig. 4H. Applying Eq. 4a to the first surface, we have

$$\frac{1.00}{\infty} + \frac{1.50}{s_1'} = \frac{1.50 - 1.00}{2} \quad \text{or} \quad \frac{1.5}{s_1'} = \frac{0.5}{2} + \frac{1}{5}$$

from which $s_1' = 30$ cm. Now, when the same equation is applied to the second surface, the following substitutions are made: $s_2 = -28$ cm, $n' = 1.50$, $n'' = 1.33$, and $r_2 = -2$ cm. Hence

$$\frac{1.50}{-28} + \frac{1.33}{s_2'} = \frac{1.33 - 1.50}{-2} \quad \text{or} \quad \frac{1.33}{s_2'} = \frac{0.17}{2} + \frac{1.50}{28}$$

This gives

$$s_2' = +9.6 \text{ cm}$$

Particular attention should be paid to the signs of the various quantities in this second step. Because the surface is concave toward the incident ray, r_2 must be assigned a negative sign. The incident rays in the glass ($n' = 1.50$) correspond to an object point M_2 which is virtual, and thus s_2 , being to the *right* of the vertex A_2 , must also be negative. The final image is formed in the water ($n'' = 1.33$) at a distance +9.6 cm from the second vertex. The positive sign of the result signifies that the image is real.

The student will find it instructive to carry through this problem by the use of the Newtonian formula (using Eqs. 4b, 4c, and 4h) and also by the method of reduced vergences (Eqs. 4j to 4l). To obtain the magnification of the image, Eq. 4f must be applied twice—once to obtain the size of the virtual object, and once to obtain the size of the final image relative to this. The final magnification is found to be $m = -1.78\times$, the symbol \times denoting “times the size of the object.”

4.8. Gaussian Formula. The basic equation 4a is of sufficient importance to warrant deriving it by two different methods, the method of wave fronts and the method of rays. Each method brings out instructive points.

In employing the wave-front method we shall have need of a well-known geometrical theorem known as the *sagitta formula*. In Fig. 4I it

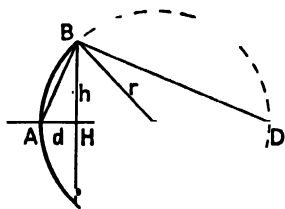


FIG. 4I. Diagram showing the geometry from which the sagitta formula may be derived.

can be shown that the triangles ABH and BDH are similar, so that

$$\frac{AH}{BH} = \frac{BII}{DII} \quad \text{or} \quad \frac{d}{h} = \frac{h}{2r - d}$$

Now, if d is very small compared with the radius r , it may be neglected in the right-hand denominator. When this is done and h is transposed to the right side, one obtains

$$d = \frac{h^2}{2r} \quad \text{SAGITTA FORMULA} \quad (4m)$$

In Fig. 4J the heavily drawn arc represents a spherical refracting surface of radius r separating two media of index n and n' . A point object at M is imaged at its conjugate point M' . A wave front BAG in the

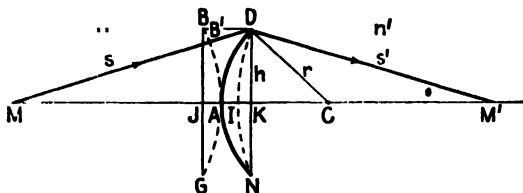


FIG. 4J. Derivation of conjugate focal relations for a single refracting surface by the method of wave fronts.

object space is shown tangent to the refracting surface, as well as the wave front DIN in the image space and tangent to that surface. By Mañus' theorem (Sec. 1.9) the optical path between these two wave fronts is the same along any ray. Therefore

$$n(B'D) = n'(AI)$$

For a paraxial ray it may be assumed that $B'D = BD$, so we have

$$(B'D) = (BD) = (JK)$$

Hence

$$n(JK) = n'(AI) \quad (4n)$$

For the two distances along the axis we may substitute $JK = JA + AK$ and $AI = AK - IK$, giving

$$n(JA + AK) = n'(AK - IK)$$

or

$$n(JA) + n(AK) = n'(AK) - n'(IK)$$

Since the small line segments in parentheses all correspond to the sagittas d of certain arcs, we may write, using the sagitta formula,

$$n \frac{h^2}{2s} + n \frac{h^2}{2r} = n' \frac{h^2}{2r} - n' \frac{h^2}{2s'}$$

This simplifies to

$$\frac{n}{s} + \frac{n}{r} = \frac{n'}{r} \quad \text{or} \quad \frac{n}{s} + \frac{n'}{s'} = \frac{n' - n}{r} \quad (4a)$$

The two approximations made in deriving this equation, namely that $B'D = BD$ and that the simplified sagitta formula is applicable, are justified for paraxial rays only. From Fig. 4J it is observed, however, that as the height h of a ray decreases, the two segments $B'D$ and BD become more nearly parallel and at the same time more nearly equal in magnitude. Furthermore the sagittas become more insignificant as compared to the radii.

Similar approximations must of course be made in deriving Eq. 4a by the ray method. Thus in Fig. 4K, if the incident and refracted rays

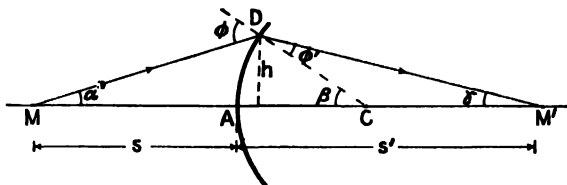


FIG. 4K. Derivation of conjugate relations for a single refracting surface by the ray method.

MD and DM' are paraxial, the angles ϕ and ϕ' will be small enough so that we may put the sines equal to the angles and use Eq. 1d:

$$\frac{\phi}{\phi'} = \frac{n'}{n} \quad (1d)$$

Now ϕ is an exterior angle in the triangle MDC and equals the sum of the opposite interior angles:

$$\phi = \alpha + \beta$$

Similarly β is an exterior angle of triangle DCM' , so that $\beta = \phi' + \gamma$, and

$$\phi' = \beta - \gamma$$

Substituting these values of ϕ and ϕ' in Eq. 1d and multiplying out, we obtain

$$n'\beta - n'\gamma = n\alpha + n\beta$$

or

$$n\alpha + n'\gamma = (n' - n)\beta$$

For paraxial rays α , β , and γ are very small angles, and we may set, $\alpha = h/s$, $\beta = h/r$, and $\gamma = h/s'$. Substituting these values and can-

celing h , we have the desired result

$$\frac{n}{s} + \frac{n'}{s'} = \frac{n' - n}{r} \quad (4a)$$

Clearly the two approximations which had to be made in this derivation are similar to those required in the wave-front method.

4.9. Derivation of Lens Makers' Formula. The thin-lens formula, Eq. 3e, may be derived from Eq. 4a by applying it to each surface independently and then adding the two equations. Let n , n' , and n'' represent the refractive indices as shown in Fig. 4H, and r_1 and r_2 the two radii. For the first surface

$$\frac{n}{s_1} + \frac{n'}{s'_1} = \frac{n' - n}{r_1} \quad (4o)$$

and for the second surface

$$\frac{n'}{s_2} + \frac{n''}{s'_2} = \frac{n'' - n'}{r_2} \quad (4p)$$

Assuming the thickness A_1A_2 of the lens to be negligible compared with the object and image distances, and noting that the image distance s'_1 becomes the object distance for the second surface but with the opposite sign, we have

$$\frac{n'}{s'_1} = -\frac{n'}{s_2}$$

We now add Eqs. 4o and 4p, obtaining

$$\frac{n}{s_1} + \frac{n''}{s'_2} = \frac{n' - n}{r_1} + \frac{n'' - n'}{r_2} \quad (4q)$$

or

$$\frac{n}{s_1} + \frac{n''}{s'_2} = \frac{n' - n}{r_1} - \frac{n' - n''}{r_2} \quad (4r)$$

If the medium on both sides of the lens is the same, $n = n''$. Furthermore s_1 and s'_2 can be called the object and image distances s and s' for the lens as a whole, so we may write

$$\frac{n}{s} + \frac{n}{s'} = \frac{n' - n}{r_1} - \frac{n' - n}{r_2}$$

or

$$\frac{n}{s} + \frac{n}{s'} = (n' - n) \left(\frac{1}{r_1} - \frac{1}{r_2} \right) \quad (4s)$$

Finally, in case the surrounding medium is air ($n = 1$), we obtain the lens makers' formula

$$\frac{1}{s} + \frac{1}{s'} = (n' - 1) \left(\frac{1}{r_1} - \frac{1}{r_2} \right) \quad (3e)$$

In the power notation of Eq. 4i, the more general formula, Eq. 4q, can be written

$$V + V' = P_1 + P_2 \quad (4t)$$

where

$$V = \frac{n}{s} \quad V' = \frac{n''}{s'} \quad (4u)$$

again using the simplification of replacing s_1 and s'_2 by s and s' , respectively. Also

$$P_1 = \frac{n' - n}{r_1} \quad \text{and} \quad P_2 = \frac{n'' - n}{r_2} \quad (4v)$$

By Eq. 4t the sum of the vergences equals the total power, so that by Eq. 4t

$$P = P_1 + P_2 \quad (4w)$$

where P is the power of the lens. Stating this in words, *the refracting power of a thin lens is equal to the sum of the powers of the two surfaces.* In terms of focal lengths, the powers are given by

$$P = \frac{n}{f} = \frac{n''}{f'}, \quad P_1 = \frac{n}{f_1} = \frac{n'}{f'_1}, \quad P_2 = \frac{n}{f_2} = \frac{n''}{f'_2} \quad (4x)$$

Problems

1. The left end of a long glass rod, of index 1.50, is ground and polished to a convex spherical surface of radius 4 cm. A small object 1 cm high is located in the air and on the axis 10 cm from the vertex. Find (a) the primary and secondary focal lengths, (b) the power of the surface, (c) the image distance, and (d) the lateral magnification.

2. The left end of a long plastic rod, of index 1.40, is ground and polished to a convex spherical surface of radius 2 cm. A small object 1 cm high is located in the air and on the axis 10 cm from the vertex. Find (a) the primary and secondary focal points, (b) the image distance, (c) the size of the image, and (d) the power of the surface.

3. Determine graphically by the oblique-ray method the image distance in Prob. 1. By the parallel-ray method, find the size of the image.

4. Determine graphically by the oblique-ray method the image distance in Prob. 2. By the parallel-ray method, find the size of the image.

5. The left end of a water trough has a transparent concave surface of radius 3 cm. A small object 2 cm high is located in the air and on the axis 12 cm from the vertex. Find (a) the primary and secondary focal points, (b) the image distance, (c) the lateral magnification and size of the image, and (d) the power of the surface.

6. Solve Prob. 2 under the assumption that the polished surface is concave instead of convex.

7. Solve Prob. 5 graphically using first the oblique-ray method and second the parallel-ray method.

8. Solve Prob. 6 graphically using first the oblique-ray method and second the parallel-ray method.

9. The left end of a long glass rod of index 1.700 is polished to a convex surface of radius 1 cm, and then submerged in clear water ($n = 1.333$). A small object 2 cm high is located in the water and on the axis 10 cm from the vertex. Calculate (a) the primary and secondary focal lengths, (b) the image distance, (c) the lateral magnification, and (d) the power of the surface.

10. A glass rod 2 cm long and of index 1.50 has both ends polished to spherical surfaces of radius 2 cm. An object 2 cm high is located on the axis 5 cm from the vertex. Find (a) the image distance for the first surface, (b) the object distance for the second surface, and (c) the final image distance. (d) What is the size of the image?

11. A parallel beam of light enters a clear plastic bead of index 1.40 and radius 2 cm. At what point beyond the bead are these rays brought to a focus?

12. A glass bead of index 1.70 and radius 1 cm is submerged in a clear liquid of index 1.30. If a parallel beam traveling in the liquid is allowed to enter the bead, at what point beyond the far side are the rays brought to a focus?

13. A hollow glass cell is made of thin glass in the form of an equiconcave lens. The radii of the two surfaces are 5 cm. When sealed airtight, this cell is submerged in water. (a) Assuming this to be a thin air lens in water, calculate its focal length, assuming the refractive index for water to be 1.333. (b) If an object is located in the water on the axis and 10 cm from the lens, locate the final image. (c) Find the power of this lens under these conditions.

14. The focal length of a thin glass lens in air is 25 cm. What will be the focal length of the same lens when it is submerged in water? Assume the refractive indices for air, water, and glass to be 1.000, 1.333, and 1.500, respectively. Calculate the power of the lens in both cases.

15. An equiconvex glass lens 3 cm thick and having radii of 4 cm is mounted in the end of a water tank. If a parallel beam of light in air enters the lens along its axis, at what point in the water will the rays come to a focus? Assume the glass to have an index 1.600 and water to have an index 1.333.

16. A small artificial flower is embedded at the center of a glass sphere of 3 cm radius. Find its apparent position and relative size, if the index is 1.50.

CHAPTER 5

THICK LENSES

When the thickness of a lens cannot be considered as small compared to its focal length, some of the thin-lens formulas of Chap. 3 are no longer applicable. The lens must be treated as a "thick lens." This term is used not only for a single homogeneous lens with two spherical surfaces separated by an appreciable distance, but also for any system of coaxial surfaces which is treated as a unit. The thick lens may therefore include several component lenses, which may or may not be in contact. We have already investigated two cases which come in this category, one of which was the combination of a pair of thin lenses spaced some distance apart (Sec. 3.12), the other the system of two spherical surfaces (Sec. 4.7). The latter represents a homogeneous thick lens. As has been shown, it is possible to determine the position of the image in this case by applying

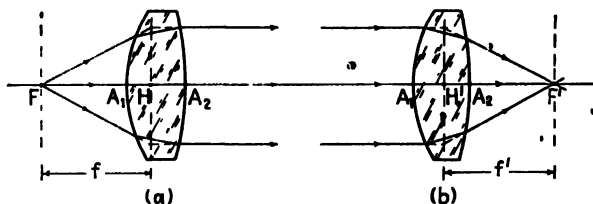


FIG. 5A. Ray diagrams showing the location of (a) the primary focal point F and focal plane, and the primary principal point H and principal plane, and (b) the secondary focal point F' and focal plane, and the secondary principal point H' and principal plane.

the equations for a single refracting surface successively to each surface. This procedure is cumbersome, however, and it would be preferable to have methods of constructing and computing the image directly from the properties of the thick lens as a whole. In the present chapter such methods are developed, based on the use of the six *cardinal points* of the lens: the two focal points, two principal points, and two nodal points. The first four of these are the most useful, and we begin by considering their location for a single homogeneous thick lens.

5.1. Focal Points and Principal Points. Ray diagrams showing the characteristics of the two focal points of a thick lens are given in Fig. 5A. In the first diagram rays diverging from the primary focal point F emerge parallel to the axis, while in the second diagram parallel incident rays are brought to a focus at the secondary focal point F' . In each case the

incident and refracted rays have been extended to their point of intersection. Transverse planes through these intersections constitute the *primary and secondary principal planes*. These planes cross the axis at points H and H' , called the *principal points*. When this graphical construction is carried out for parallel rays at different heights above or below the axis, the positions of the principal points and focal points will be found to vary somewhat. This implies that the focal planes and principal planes are in reality curved surfaces. For paraxial rays, however, all four points F , F' , H , and H' are to be regarded as fixed, and the corresponding "planes" as truly plane.

If the medium is the same on both sides of the lens, the primary focal length f is exactly equal to the secondary focal length f' . For this to be true the focal lengths must, as is shown in the figure, be measured from the focal points to their respective principal points H and H' and *not* to their respective vertices A_1 and A_2 . In general the principal points and focal points are not located symmetrically with respect to the lens but are at different distances from the vertices. As a lens of a given focal

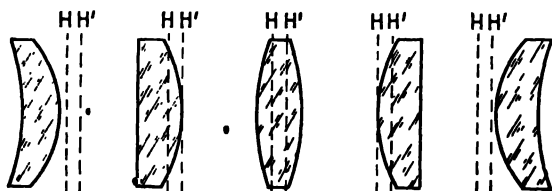


FIG. 5B. Illustrating the variation of the positions of the primary and secondary principal planes, as a thick lens of a given focal length is subject to "bending."

length is "bent" (Fig. 5B), departing in either direction from the symmetrical shape of an equiconvex lens, the principal points shift in the opposite direction. For the equiconvex lens shown in the center, H and H' divide the lens thickness into three approximately equal parts. In the plano-convex lens one principal point falls exactly at the vertex of the convex surface, and the other falls about one-third of the thickness inside this vertex. For meniscus lenses of considerable thickness and curvature, H and H' may lie completely outside the lens.

5.2. Conjugate Relations. In order to trace any ray through a thick lens, the positions of the focal points and principal points must first be determined. Once this has been done, either graphically or by computation, the parallel-ray construction can be used to locate the image as shown in Fig. 5C. The construction procedure follows that given in Fig. 3F for a thin lens, except that here all rays in the region between the two principal planes are drawn parallel to the axis. This requirement results from the fact that by definition the primary and secondary

principal planes are conjugate planes for which the lateral magnification is unity and has a positive sign. Any ray starting from one of these planes must, after emerging from the lens, appear to come from that point on the other plane which is at the same distance from the axis.

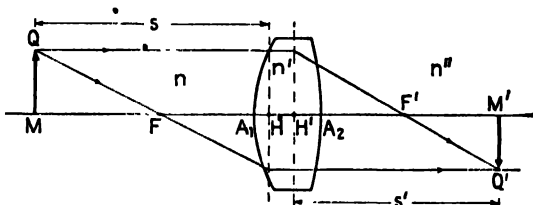


FIG. 5C. Parallel-ray method for the graphical location of an image formed by a thick lens.

By comparison of Fig. 5C with Fig. 3F' and with the derivation of Sec. 3.13, it will be found that *provided the object and image distances are measured from the principal points* the geometry of the right triangles is the same as that for thin lenses, and we may apply to the thick lens the Gaussian lens formula, namely

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \quad (5a)$$

In the more general case where the media are different on the two sides of the lens, the equation corresponding to 5a is

$$\frac{n}{s} + \frac{n'}{s'} = \frac{n}{f} = \frac{n'}{f'} \quad (5b)$$

or

$$V + V' = P \quad (5c)$$

In this general case as well, the focal lengths and the object and image distances must be measured from the principal planes and not from the

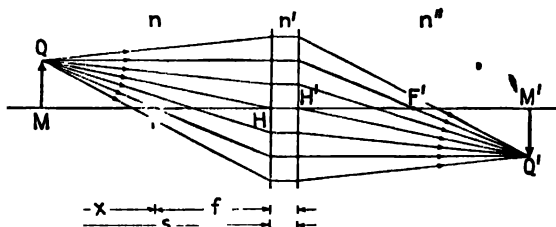


FIG. 5D. Principal planes are planes of unit magnification.

lens surfaces. Figure 5D emphasizes this fact by showing that for the purposes of graphical construction the lens may be regarded as replaced

by its two principal planes. This figure also illustrates the unit magnification that exists between the principal planes in that any ray intersects these two planes at the same ordinate.

5.3. Graphical Construction. The oblique-ray method of construction may be used to find graphically the focal points and principal points of a thick lens. As an illustration consider a glass lens of index 1.50, thickness 2 cm, and radii $r_1 = +3$ cm, $r_2 = -5$ cm.⁴ By graphical construction or by formula the focal points for each surface alone are first obtained and then marked on the axis of the lens as shown in Fig. 5E. All known distances are tabulated as follows:

$$\begin{array}{llll} r_1 = +3 \text{ cm} & d = 2 \text{ cm} & f_1 = +6 \text{ cm} & f_2 = +15 \text{ cm} \\ r_2 = -5 \text{ cm} & n = 1.50 & f'_1 = +9 \text{ cm} & f'_2 = +10 \text{ cm} \end{array}$$

A parallel ray 4 is refracted by the first surface toward the secondary focal point F'_1 of this surface; within the lens this ray has the direction labeled 5. Line 7 is next drawn through C_2 parallel to ray 5, extending

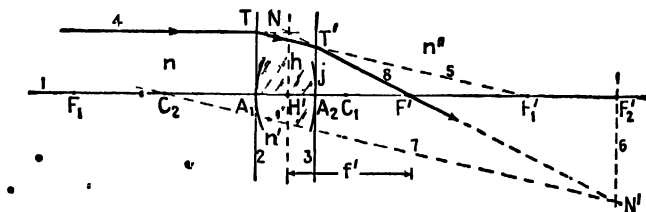


FIG. 5E. Oblique-ray construction for locating the secondary focal point and secondary principal point of a thick lens.

it to its intersection with the focal plane of the second surface at N' . Finally ray 8 is drawn from T' through N' . The intersection of ray 8 with the lens axis locates the secondary focal point F'' of the lens, while its intersection with the incident ray 4 locates the secondary principal plane H' .

By turning the lens around and repeating this procedure, the positions of the primary focal point F and of the primary principal point H can be determined. The student will find it well worth while to carry out this construction and to check the results by measuring the focal lengths to verify the fact that they are equal. It is to be noted that, in accordance with the assumption of paraxial rays, any refraction is taken as occurring at the plane tangent to the boundary at its vertex.

5.4. Thick-lens Formulas. The focal lengths and power of a one-component thick lens, as well as the positions of the focal points and principal points, can be calculated by the use of the following equations:

Gaussian formulas

$$\frac{n}{f} = \frac{n}{f_1} + \frac{n'}{f_2} - c \frac{n n'}{f_1 f_2}$$

$$\frac{n}{f_1} = \frac{n' - n}{r_1} = \frac{n'}{f'_1}$$

$$\frac{n'}{f_2} = \frac{n'' - n'}{r_2} = \frac{n''}{f'_2}$$

$$\frac{n}{f} = \frac{n''}{f'}$$

$$c = \frac{A_1 A_2}{n'} = \frac{d}{n'}$$

$$A_1 F = -f \left(1 - n' \frac{c}{f_2} \right)$$

$$A_1 H = +f n' \frac{c}{f_2}$$

$$A_2 F' = +f' \left(1 - n' \frac{c}{f_1} \right)$$

$$A_2 H' = -f' n' \frac{c}{f_1}$$

Power formulas

$$P = P_1 + P_2 - c P_1 P_2 \quad (5d)$$

$$P_1 = \frac{n}{f_1} = \frac{n'}{f'_1} = \frac{n' - n}{r_1} \quad (5e)$$

$$P_2 = \frac{n'}{f_2} = \frac{n''}{f'_2} = \frac{n'' - n'}{r_2} \quad (5f)$$

$$P = \frac{n}{f} = \frac{n''}{f'} \quad (5g)$$

$$c = \frac{A_1 A_2}{n'} = \frac{d}{n'} \quad (5h)$$

$$A_1 F = -\frac{n}{P} (1 - c P_2) \quad (5i)$$

$$A_1 H = +\frac{n}{P} c P_2 \quad (5j)$$

$$A_2 F' = +\frac{n''}{P'} (1 - c P_1) \quad (5k)$$

$$A_2 H' = -\frac{n''}{P} c P_1 \quad (5l)$$

The subscripts 1 and 2 refer to the first and second surfaces of the lens, respectively. These equations are derived from geometrical relations that may be obtained* from a diagram like Fig. 5E.

In the design of certain optical systems it is convenient to know the *vertex power* of a lens. This power, sometimes called the *effective power*, is given as

$$P_v = \frac{P_1}{1 - c P_1} + P_2 \quad (5m)$$

and is defined as the reciprocal of the distance from the back surface of the lens to the secondary focal point. This distance is commonly called the *back focal length*. Equation 5m is derived by substituting the value of P from Eq. 5d in Eq. 5k and solving for $1/A_2 F'$, which is the vertex power P_v . In this derivation the lens is assumed to be in air so that $n'' = 1$.

In a similar way the distance from the primary focal point to the front surface of a lens is called the *front focal length*, and the reciprocal of this distance is called the *neutralizing power*. The name is derived from the fact that a thin lens of this specified power will, upon contact with the front surface, give zero power to the combination.

* Derivations of these equations are left as exercises for the student. See Probs. 17 and 18 at the end of this chapter.

To illustrate the use of the above formulas, consider again the lens discussed in the preceding section. The quantities given are $r_1 = +3$ cm, $r_2 = -5$ cm, $n = n'' = 1$, $n' = 1.50$, and $d = 2$ cm.

Applying the set of Gaussian formulas, Eqs. 5e, 5f, and 5g, we obtain for the individual focal lengths of the surfaces

$$\begin{aligned}\frac{n}{f_1} &= \frac{n' - n}{r_1} = \frac{1.50 - 1.00}{3} = 0.167, & f_1 &= \frac{1.00}{0.167} = +6 \text{ cm}, \\ & & f_1' &= \frac{3}{0.5} 1.5 = +9 \text{ cm} \\ \frac{n'}{f_2} &= \frac{n'' - n'}{r_2} = \frac{1.00 - 1.50}{-5} = 0.100, & f_2 &= \frac{1.50}{0.100} = +15 \text{ cm}, \\ & & f_2' &= \frac{1.0}{0.1} = +10 \text{ cm} \\ c &= \frac{d}{n'} = \frac{2}{1.50} = 1.333 \text{ cm}\end{aligned}$$

The focal length of the lens is calculated from Eq. 5d, which gives

$$\frac{n}{f} = 0.167 + 0.100 - 1.333 \times 0.167 \times 0.100 = 0.245$$

Since air exists on both sides of the lens, $n = n'' = 1$, and the primary and secondary focal lengths are equal:

$$f = f' = \frac{1.00}{0.245} = 4.08 \text{ cm}$$

The principal points and focal points are located by the use of Eqs. 5i, 5j, 5k, and 5l, as follows:

$$\begin{aligned}A_1F &= -4.08 \left(1 - 1.50 \frac{1.333}{15} \right) = -4.08 \times 0.867 = -3.540 \text{ cm} \\ A_1H &= +4.08 \times 1.50 \frac{1.333}{15} = +4.08 \times 0.133 = +0.542 \text{ cm} \\ A_2F' &= +4.08 \left(1 - 1.50 \frac{1.333}{9} \right) = +4.08 \times 0.778 = +3.175 \text{ cm} \\ A_2H' &= -4.08 \times 1.50 \frac{1.333}{9} = -4.08 \times 0.222 = -0.906 \text{ cm}\end{aligned}$$

Positive signs represent distances measured to the right of the reference point and negative signs those measured to the left of it. By adding the magnitudes of the first two intervals A_1F and A_1H , the primary focal length $FH = 4.08$ cm is obtained, and this serves as a check upon the

calculations. Similarly the addition of the two intervals A_2F' and A_2H' gives the secondary focal length $H'F'$ as 4.08 cm also.

Once the principal points and focal points of a thick lens have been located, the thin-lens formulas may be applied to find object and image distances, through the relations

$$\frac{n}{s} + \frac{n''}{s'} = \frac{n}{f} \quad \text{or} \quad V + V' = P \quad (5b \text{ and } c)$$

and the magnification through Eqs. 3c or 3i, in case the medium is the same on both sides. The magnification when the medium is different will be discussed in the next section. In using these formulas it is important that all distances must be measured from the principal points and not from the lens surfaces or from the center of the lens.

5.5. Nodal Points and Optical Center.

Of all the rays that pass through a lens from an off-axis object point to its corresponding image point, there will always be one for which the direction of the ray in the image space is the same as that in the object space, *i.e.*, the segments of the ray before reaching the lens, and after leaving it, are parallel. The two points at which these segments, if projected, intersect the axis are called the *nodal points*, and the transverse planes through them are called the *nodal planes*. This third pair of cardinal points and their associated planes are shown in Fig. 5F, which also shows the optical center of the lens at C . It is readily shown that if the medium on both sides is the same, the nodal points N and N' coincide with the principal points H and H' , but that if the two media have different indices, the principal points and the nodal points will be separate. Since the incident and emergent rays make equal angles with the axis, the nodal points are called conjugate points of *unit positive angular magnification*.

If the ray is to emerge parallel to its original direction, the two surface elements of the lens where it enters and leaves must be parallel to each other so that the effect is like that of a plane-parallel plate. A line between these two points crosses the axis at the *optical center* C . It is therefore through the optical center that the undeviated ray must be drawn in all cases. It has the interesting property that its position, depending as it does only on the radii of curvature and thickness of the

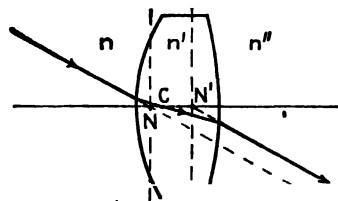


FIG. 5F. Illustrating the properties and positions of the nodal points, nodal planes, and optical center of a thick lens.

lens, does not vary with color of the light. All the six cardinal points will in general have a slightly different position for each color.

Figure 5*G* will help to clarify the different significance of the nodal points and the principal points. This figure is drawn for $n'' \neq n$, so that the two sets of points are separate. Ray 11 through the secondary nodal point is parallel to ray 10, the latter being incident in the direction of the primary nodal point N . On the other hand both these segments intersect the principal planes at the same distance above the principal points H and H' . From the small parallelogram at the center of the

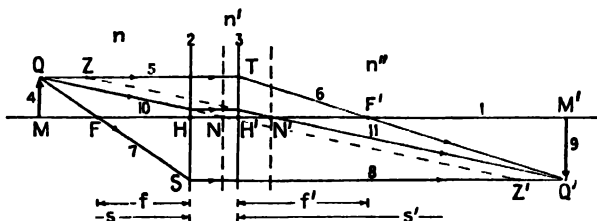


FIG. 5*G*. Parallel-ray method applied to the graphical location of nodal points and planes.

diagram, it is observed that the distance between nodal planes is exactly equal to the distance between principal planes. In general, therefore,

$$NN' = HH' \quad (5n)$$

Furthermore in this case, where the initial and final values of the refractive index differ, the focal lengths, which are measured from the principal points, are no longer equal. The primary focal length FH is equal to the distance $N'F'$, while the secondary focal length $H'F'$ is equal to FN :

$$f = FH = N'F' \quad \text{and} \quad f' = H'F' = FN \quad (5o)$$

Nodal points may be determined graphically, as shown in Fig. 5*G*, by measuring off the distance $ZQ = HH' = Z'Q'$ and drawing straight lines through QZ' and ZQ' . From the geometry of this diagram, the lateral magnification y'/y is given by

$$m = \frac{y'}{y} = -\frac{s'}{s} + \frac{HN}{HN'}$$

where

$$HN = f' \frac{n'' - n}{n''}$$

When the object and image distances s and s' are, as usual, measured from their corresponding principal points H and H' , Eq. 5*b* is also valid in this case for paraxial rays.

5.6. Other Cardinal Points. In thick-lens problems a knowledge of the six cardinal points, comprising the focal points, principal points, and nodal points, is always adequate to obtain solutions. Other points of lesser importance but still of some interest are (1) negative principal points, and (2) negative nodal points. *Negative principal points* are conjugate points for which the lateral magnification is unity and negative. They lie at twice the focal length and on opposite sides of the lens. *Negative nodal points* lie as far from the focal points as the ordinary cardinal nodal points, but on opposite sides. Their position is such that the angular magnification is unity and negative. Although a knowledge of these two pairs of cardinal points is not essential to the solution of optical problems, in certain cases considerable simplification is achieved by using them.

5.7. Thin-lens Combination as a Thick Lens. A combination of two or more thin lenses may also be referred to as a thick lens. This is because of the fact that the optical properties of a set of coaxially mounted

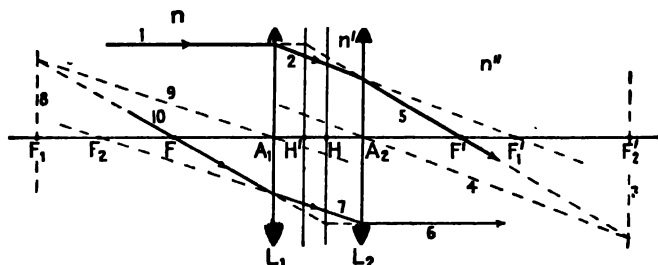


FIG. 5H. Foci and principal points of a combination of two positive thin lenses.

lenses can be conveniently treated in terms of only two focal points and two principal points. If the object space and image space have the same refractive index (and this is nearly always the case), the nodal points and planes coincide with the principal points and planes.

A combination of two thin lenses with focal lengths of 8 and 9 cm respectively is shown in Fig. 5H. By the oblique-ray method the focal points F and F' and the principal points H and H' have been determined graphically. In doing so the refraction at each lens was considered in the same way as the refraction at the individual surfaces of the thick lens of Fig. 5E. There is a strong resemblance between these two diagrams; *i.e.*, for a thin lens we assume that all of the deviation occurs at one plane, just as for a single surface. This assumption is justified only when the separation of the principal planes of the lens can be neglected. The definition of a thin lens is just a statement of this fact: *a thin lens is one in which the two principal planes and the optical center*

coincide at the geometrical center of the lens. The locations of the centers of the two lenses in this example are labeled A_1 and A_2 in Fig. 5H.

A comparison of Fig. 5H with Fig. 5D will show that the order of the principal points has been reversed, i.e., they are *crossed* in Fig. 5H. The parallel-ray method of construction of object and image rays for such a case is shown in Fig. 5I. All object rays are terminated at the primary principal plane H , where they are projected backward with unit magni-

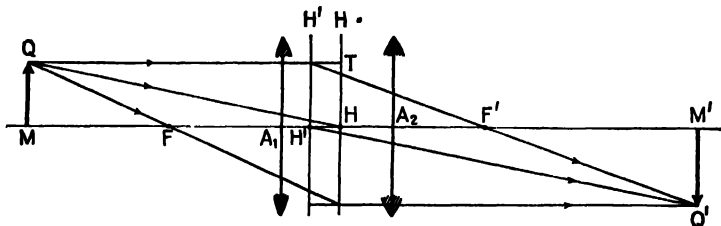


FIG. 5I. Parallel-ray construction when the principal points are crossed.

cation to the secondary principal plane H' . From there the image rays are drawn following the principles of the parallel-ray method described earlier.

A diagram for a combination of a positive and a negative lens is given in Fig. 5J. The construction lines are not shown, but the graphical procedure used in determining the paths of the two rays is the same as that shown in Fig. 5II. Note here that the final principal points H and H'

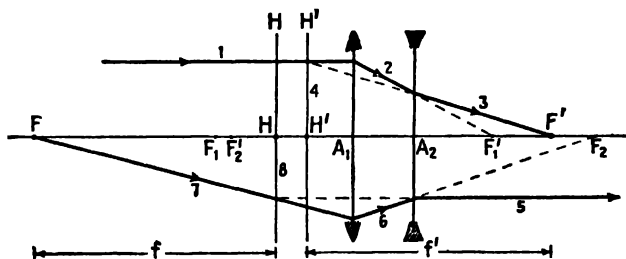


FIG. 5J. Showing the oblique-ray method applied to positive and negative thin lenses to find the focal points and principal points.

lie outside the interlens space but that the focal lengths f and f' measured from these points are as usual equal. The lower ray, although shown traveling from left to right, is graphically constructed by drawing it from right to left.

The positions of the cardinal points of a combination of two thin lenses in air can be calculated by means of the thick-lens formulas given in Sec. 5.4. As used for thin lenses in place of individual refracting surfaces,

A_1 and A_2 become the two lens centers, while f_1 , f_2 and P_1 , P_2 become their separate focal lengths and powers respectively. The latter are given by

$$P_1 = \frac{n_1 - n}{r_1} + \frac{n' - n_1}{r'_1} = \frac{n}{f_1} \quad P_2 = \frac{n_2 - n'}{r_2} + \frac{n'' - n_2}{r'_2} = \frac{n'}{f_2} \quad (5p)$$

where r_1 and r'_1 are the radii of the first lens of index n_1 , and r_2 and r'_2 are the radii of the second lens of index n_2 . The surrounding media have indices n , n' , and n'' (see Fig. 5H). The other formulas, Eqs. 5d, 5g, 5h, 5i, 5j, 5k, and 5l, remain unchanged.

To illustrate the use of these formulas, let us consider the following problem on a lens combination similar to that shown in Fig. 5J:

An equiconvex lens with radii of 4 cm and index $n_1 = 1.50$ is located 2 cm in front of an equiconcave lens with radii of 6 cm and index $n_2 = 1.60$. The lenses are to be considered as thin. The surrounding media have indices $n = 1.00$, $n' = 1.33$, and $n'' = 1.00$. Find (a) the power, (b) the focal lengths, (c) the focal points, and (d) the principal points of the system.

In this instance we shall solve the problem by the use of the power formulas. By Eqs. 5p the powers of the two lenses in their surrounding media are

$$P_1 = \frac{1.50 - 1.00}{0.04} + \frac{1.33 - 1.50}{-0.04} = 12.50 + 4.17 = +16.67 \text{ D}$$

$$P_2 = \frac{1.60 - 1.33}{-0.06} + \frac{1.00 - 1.60}{0.06} = -4.45 - 10.0 = -14.45 \text{ D}$$

From Eq. 5h the reduced interval c is

$$c = \frac{0.02}{1.33} = 0.015$$

By Eq. 5d,

$$P = 16.67 - 14.45 + 0.015 \times 16.67 \times 14.45$$

or

$$P = +5.84 \text{ D} \quad \text{Ans. (a)}$$

Using Eq. 5g,

$$\left. \begin{aligned} f &= \frac{n}{P} = \frac{1.00}{5.84} = 0.171 \text{ m} = 17.1 \text{ cm} \\ f' &= \frac{n''}{P} = \frac{1.00}{5.84} = 0.171 \text{ m} = 17.1 \text{ cm} \end{aligned} \right\} \text{ Ans. (b)}$$

By Eqs. 5i, 5j, 5k, and 5l,

$$A_1F = -\frac{1.00}{5.84}(1 + 0.015 \times 14.45) = -0.208 \text{ m} = -20.8 \text{ cm}$$

$$A_1H = +\frac{1.00}{5.84}0.015(-14.45) = -0.037 \text{ m} = -3.7 \text{ cm} \quad \text{Ans. (c)}$$

$$A_2F' = +\frac{1.00}{5.84}(1 - 0.015 \times 16.67) = +0.128 \text{ m} = +12.8 \text{ cm} \quad \text{Ans. (d)}$$

$$A_2H' = -\frac{1.00}{5.84}0.015 \times 16.67 = -0.043 \text{ m} = -4.3 \text{ cm}$$

As a check on these results we find that the difference between the first two intervals A_1F and A_1H gives the primary focal length $FH = 17.1 \text{ cm}$. Similarly the sum of the second two intervals A_2F' and A_2H' gives the secondary focal length $H'F' = 17.1 \text{ cm}$.

5.8. Thick-lens Combinations. The problem of calculating the positions of the cardinal points of a thick lens consisting of a combination of several component lenses of appreciable thickness is one of considerable difficulty, but one which may be solved by use of the principles already given. If, in a combination of two lenses such as that in Fig. 5H, the individual lenses cannot be considered as thin, each must be represented by a pair of principal planes. There are thus two pairs of principal points, H_1 and H_1' for the first lens and H_2 and H_2' for the second, and the problem is to combine these to find a single pair H and H' for the combination, and to determine the focal lengths. By carrying out a construction similar to Fig. 5E for each lens separately, it is possible to locate the principal points and focal points of each. Then the construction of Fig. 5H may be accomplished, taking account of the unit magnification between principal planes.

Formulas may be given for the analytical solution of this problem, but because of their complexity they will not be given here.* Instead, we shall describe a method of determining the positions of the cardinal points of any thick lens by direct experiment.

5.9. Nodal Slide. The nodal points of a single lens, or of a combination of lenses, may be located experimentally by mounting the system on a nodal slide. This is merely a horizontal support which permits rotation of the lens about any desired point on its axis. As is shown in Fig. 5K, light from a source S is sent through a slit Q , adjusted to lie at the secondary focal point of the lens. Emerging as a parallel beam, this

* These equations are given for example in G. S. Monk, "Light, Principles and Experiments," 1st ed., McGraw-Hill Book Company, Inc., New York,

light is reflected back on itself by a fixed plane mirror M , passing again through the lens system and being brought to a focus at Q' . This image of the slit is formed slightly to one side of the slit itself on the white face of one of the slit jaws. The nodal slide carrying the lens system is now rotated back and forth and the lens repeatedly shifted, until the rotation produces no motion of the image Q' . When this condition is reached, the axis of rotation N' locates one nodal point. By turning the nodal slide end-for-end and repeating the process, the other nodal point N is found. When performed in air, this experiment of course locates the

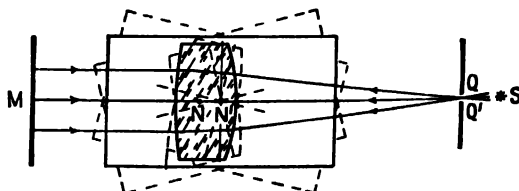


FIG. 5K. Illustrating the use of a nodal slide in locating nodal points.

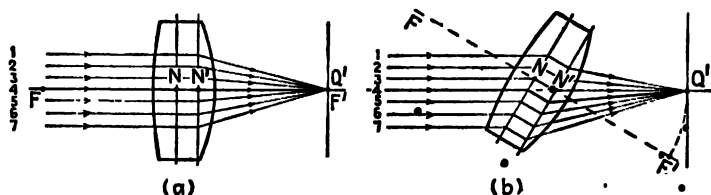


FIG. 5L. Rotation about a nodal point shifts the rays but not the image point.

principal points as well, and the distance $N'Q'$ is an accurate measure of the focal length.

The principle of this method of rotation about a nodal point is illustrated in Fig. 5L. In the first diagram ray 4 along the axis passes through N and N' to the focus at Q' . In the second diagram the lens system has been rotated about N' and the same bundle of rays passes through the lens, coming to a focus at the same point Q' . Ray 3 is now directed towards N and ray 4 towards N' . When projected across from the plane of N to that of N' , the rays still converge towards Q' even though F' is now shifted to one side. Note that ray 3 approaches N in exactly the same direction that it leaves N' , corresponding to the defining condition for the nodal points.

If a camera lens is pivoted about its secondary nodal point and a long strip of photographic film is curved to a circular arc of radius f' , a continuous picture covering a very wide angle may be taken. Such an

instrument, shown schematically in Fig. 5M, is known as a *panoramic camera*. The shutter usually consists of a vertical slit just in front of the film, which moves with the rotation so that it always remains centered on the lens axis.

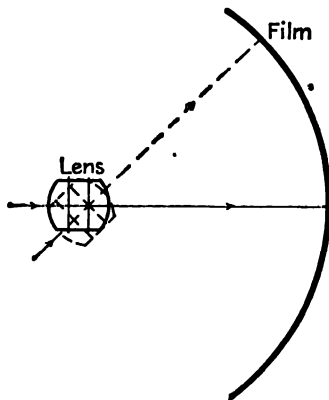


FIG. 5M. In the panoramic camera the lens rotates about a nodal point as a center.

Problems

1. A plano-convex lens 3 cm thick is made of glass of index 1.50. If the first surface has a radius of 3 cm, find (a) the focal length of the lens, (b) the power P , and (c) the distance from the first vertex to the two principal planes.
2. An equiconvex lens has an index of 1.50, radii of 5 cm, and a thickness of 1 cm. Calculate (a) the power P , (b) the focal length f , and (c) the distances from the first vertex to the two principal planes.
3. A lens with radii $r_1 = +2$ cm and $r_2 = +4$ cm has a thickness of 1 cm and an index of 1.50. Calculate (a) the power P , (b) the focal length f , and (c) the distances from the first vertex to the two principal planes.
4. A lens with radii $r_1 = +4$ cm and $r_2 = +2$ cm has a thickness of 1 cm and an index of 1.50. Calculate (a) the power P , (b) the focal length f , and (c) the distances from the first vertex to the two principal planes.
5. Solve Prob. 1 graphically, locating the focal points and principal points.
6. Solve Prob. 2 graphically, locating the focal points and principal points.
7. Solve Prob. 3 graphically, locating the focal points and principal points.
8. Solve Prob. 4 graphically, locating the focal points and principal points.
9. Two thin equiconvex lenses with radii of curvature 24 cm and index 1.60 are located 4 cm apart. Find the positions of the focal points and principal points. What is the combined power of the system? Make a diagram to scale.
10. A thin double-convex lens of 5 cm focal length is situated 2.5 cm in front of a thin double-concave lens of 7 cm focal length. Calculate for the combination, (a) the focal length, (b) the positions of the principal planes, and (c) the positions of the focal planes.
11. A thin double-concave lens of 6 cm focal length is located 2 cm in front of a thin double-convex lens of 4 cm focal length. Calculate for the combination (a) the focal length, (b) the positions of the principal planes, and (c) the positions of the focal planes.

12. Three thin lenses having powers of $+10\text{ D}$, -10 D , and $+10\text{ D}$ respectively are placed with their axes in the same straight line and 3 cm apart. (a) Graphically locate the secondary focal plane and its corresponding principal plane. (b) Measure the focal length and compare it with the focal length obtained with the three lenses in contact.

13. Two thin lenses with powers of $+10\text{ D}$ and -10 D respectively are placed with their axes in the same straight line and 5 cm apart. (a) Locate graphically the secondary focal plane and its corresponding principal plane. (b) Measure the focal length and compare it with the focal length of the two lenses in contact.

14. A convex glass lens 3 cm thick, with radii $r_1 = +2\text{ cm}$ and $r_2 = -1\text{ cm}$, forms one end of a water trough. (a) Calculate the focal lengths of the two surfaces separately. (b) Locate graphically the second focal point of the system. (c) Locate graphically the secondary principal point. (d) Determine the answer to (b) by the use of the appropriate equations. (e) From the right triangles in the diagram, calculate the position of H' .

15. Solve Prob. 14 for the primary focal point and principal point.

16. Two thin lenses have powers of $+10\text{ D}$ and -10 D respectively. What must be their separation if the combination is to have a power of $+5\text{ D}$? Locate both focal points and both principal planes.

17. From similar triangles in the geometry of Fig. 5E, derive Eq. 5k. (NOTE: Pairs of similar triangles give the required proportionalities.)

18. From similar triangles in Fig. 5E, derive Eq. 5l.

19. A convex glass lens 3 cm thick and having radii $r_1 = +3\text{ cm}$ and $r_2 = -1\text{ cm}$ is located in one end of a water tank. (a) Calculate the focal lengths of the two surfaces separately. (b) Find the focal points of the combination. (c) Find the principal points. (d) Find the nodal points. Assume indices 1.00, 1.50, and 1.33 for air, glass, and water respectively.

20. After calculating the answer to (a) in Prob. 19, find the answers to (b), (c), and (d) graphically.

21. An object 2 cm high is located 15 cm in front of the convex surface of the lens in Prob. 1. Find (a) the distance from the second surface to the image, and (b) the image size. Solve graphically and check your answer by formula.

22. An object 3 cm high is located on the axis and 23.5 cm from the first principal point of the lens combination in Prob. 9. Find (a) the distance from the second lens to the image, and (b) the image size.

23. An object 5 cm high is located on the axis and 25 cm in front of the first lens of the combination in Prob. 13. Find (a) the distance from the second lens to the image, and (b) the image size.

24. From Eqs. 5d and 5i, derive an equation like Eq. 5m but for the neutralizing power P_n of a thick lens.

CHAPTER 6

SPHERICAL MIRRORS

A spherical reflecting surface has image-forming properties similar to those of a thin lens or of a single refracting surface. The image from a spherical mirror is in some respects superior to that from a lens, notably in the absence of chromatic effects due to dispersion that always accompany the refraction of white light. Therefore mirrors are occasionally used in place of lenses in optical instruments, but their applications are not so broad as those of lenses because they do not offer the same possibilities for correction of the other aberrations of the image (Chap. 9).

Because of the simplicity of the law of reflection as compared to the law of refraction, the quantitative study of image formation by mirrors is easier than in the case of lenses. Many features are the same, and these we shall pass over rapidly, putting the chief emphasis upon those characteristics which are different. To begin with, we restrict the discussion to images formed by paraxial rays.

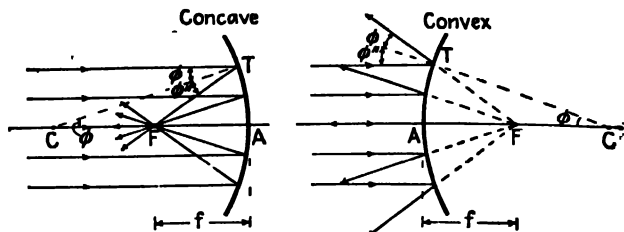


FIG. 6A. The primary and secondary focal points of spherical mirrors coincide.

6.1. Focal Point and Focal Length. Diagrams showing the reflection of a parallel beam of light by a concave mirror and by a convex one are given in Fig. 6A. A ray striking the mirror at some point such as T obeys the law of reflection $\phi'' = \phi$. All rays are shown as brought to a common focus at F , although this will be strictly true only for paraxial rays. The point F is called the *focal point* and the distance FA the *focal length*. In the second diagram the reflected rays diverge as though they came from a common point F . Since the angle TCA also equals ϕ , the triangle TCF is an isosceles one, and in general $CF = FT$. But for very small angles ϕ (paraxial rays), FT approaches equality with FA . Hence

$$(FA) = \frac{1}{2}(CA) \quad \text{or} \quad f = -\frac{1}{2}r \quad (6a)$$

and the focal length equals one-half the radius of curvature (see also Eq. 6d).

The negative sign is introduced in Eq. 6a so that the focal length of a concave mirror, which behaves like a positive or converging lens, will also be positive. According to the sign convention of Sec. 3.10, the radius of curvature is negative in this case. The focal length of a convex mirror, which has a positive radius, will then come out to be negative. This sign convention is chosen as being consistent with that used for lenses; it gives converging properties to a mirror with positive f and diverging properties to a mirror with negative f . By the principle of reversibility it may be seen from Fig. 6A that the primary and secondary focal points of a mirror coincide. In other words, it has but one focal point.

As before, a transverse plane through the focal point is called the focal plane.

Its properties, as shown in Fig. 6B, are similar to those of either focal plane of a lens; for example, parallel rays incident at any angle with the optic axis are brought to a focus at some point in the focal plane. The image Q' of a distant off-axis point object occurs at the intersection with the focal plane of that ray which goes through the center of curvature C .

6.2. Graphical Constructions. Figure 6C, which illustrates the formation of a real image by a concave mirror, is self-explanatory. When the

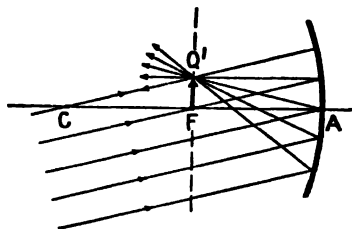


FIG. 6B. Parallel rays incident at an angle with the axis of a concave mirror are brought to a focus in the focal plane.

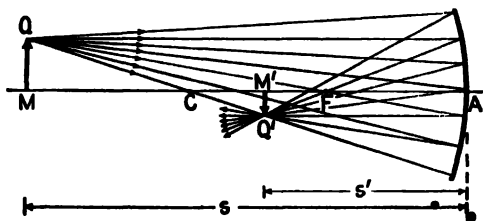


FIG. 6C. Real image due to a concave mirror.

object MQ is moved toward the center of curvature C , the image also approaches C and increases in size until when it reaches C it is the same size as the object. The conditions when the object is between C and F may be deduced from the interchangeability of object and image as applied to this diagram. When the object is inside the focal point, the image is virtual as in the case of a converging lens. The methods of graphically constructing the image follow the same principles as were

used for lenses, including the fact that paraxial rays must be represented as deflected at the tangent plane instead of at the actual surface.

An interesting experiment can be performed with a large concave mirror set up under the condition of unit magnification, as shown in Fig. 6D. A bouquet of flowers is suspended upside down in a box and illuminated by a shaded lamp S . The large mirror is placed with its center of curvature C at the top surface of the stand, on which a real vase is placed. The

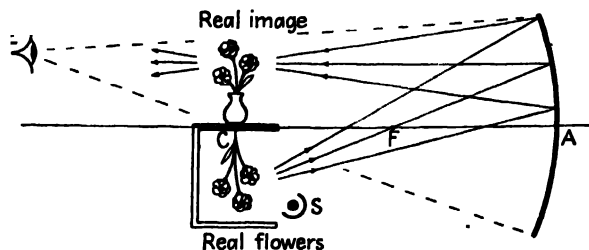


Fig. 6D. Illusion produced by a real image of unit magnification.

observer's eye at E sees a perfect reproduction of the bouquet, not merely as a picture but as a faithful three-dimensional replica, which creates a strong illusion that it is a real object. As shown in the diagram, the rays diverge from points on the image just as they would were the real object in the same position.

The parallel-ray method of construction is given for the case of a concave mirror in Fig. 6E. Three rays leaving Q are, after reflection, brought

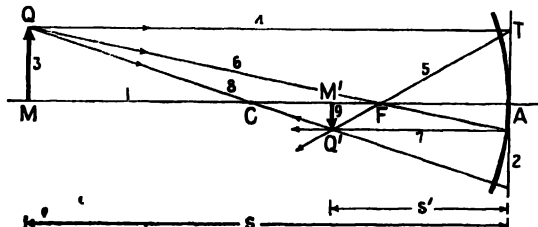


Fig. 6E. Parallel-ray method for graphically locating the image formed by a concave mirror.

to the conjugate point Q' . The image is real, inverted, and smaller than the object. Ray 4, drawn parallel to the axis is, by definition of the focal point, reflected through F . Ray 6 drawn through F is reflected parallel to the axis, and ray 8 through the center of curvature strikes the mirror normally and is reflected back on itself. The crossing point of any two of these rays is sufficient to locate the image.

A similar procedure is applied to a convex mirror in Fig. 6F. The rays from the object point Q , after reflection, diverge from the conjugate point Q' . Ray 4, starting parallel to the axis, is reflected as if it came from F . Ray 6 toward the center of curvature C is reflected back on itself, while ray 7 going toward F is reflected parallel to the axis. Since the rays never pass through Q' , the image $Q'M'$ in this case is virtual.

The oblique-ray method may also be used for mirrors, as is illustrated in Fig. 6G for a concave mirror. After drawing the axis 1 and the mirror 2, we lay out the points C and F and draw a ray 3 making any arbitrary angle with the axis. Through F , the broken line 4 is then drawn parallel to 3. Where this line intersects the mirror at S , a parallel ray 6 is drawn backward to intersect the focal plane at P . Ray 7 is then drawn

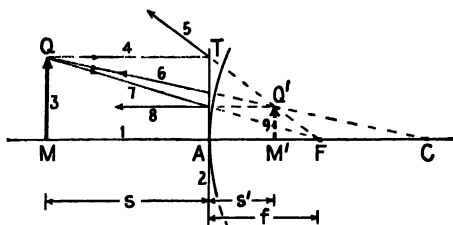


FIG. 6F. Parallel-ray method for graphically locating the image formed by a convex mirror.

through TP and intersects the axis at M' . By this construction M and M' are conjugate points, and 3 and 7 are the parts of the ray in object and image spaces. The principle involved in this construction is obvious from the fact that if 3 and 4 were parallel incident rays they would come to a focus at P in the focal plane. If in place of ray 4 another ray were drawn through C and parallel to ray 3, it too would cross the focal plane at P . A ray through the center of curvature would be reflected directly back upon itself.

6.3. Mirror Formulas. In order to be able to apply the standard lens formulas of the preceding chapters to spherical mirrors with as little change as possible, we must adhere to the following sign conventions:

1. Distances measured from left to right are positive, while those measured from right to left are negative.
2. Incident rays travel from left to right and reflected rays from right to left.
3. The focal length is measured from the focal point to the vertex. This gives f a positive sign for concave mirrors and a negative sign for convex mirrors.

4. *The radius is measured from the vertex to the center of curvature.* This makes r negative for concave mirrors and positive for convex mirrors.
5. *Object distances s and image distances s' are measured from the object and from the image respectively to the vertex.* This makes both s and s' positive and the object and image real when they lie to the left of the vertex, while they are negative and virtual when they lie to the right.

The last of these sign conventions implies that for mirrors the object space and the image space coincide completely, the actual rays of light always lying in the space to the left of the mirror. Since the refractive index of the image space is the same as that of the object space, the n' of the previous equations becomes numerically equal to n .

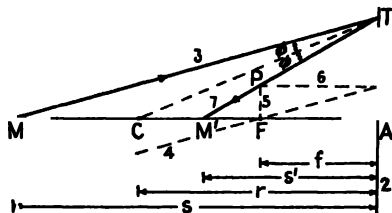


FIG. 6G. Oblique-ray method for locating the image formed by a concave mirror.

The following is a simple derivation of the formula giving the conjugate relations for a mirror. Referring to Fig. 6G it is observed that by the law of reflection the radius CT bisects the angle MTM' . Using a well-known geometrical theorem, we may then write the proportion

$$\frac{MC}{MT} = \frac{CM'}{M'T}$$

Now, for paraxial rays, $MT \cong MA = s$ and $M'T \cong M'A = s'$, where the symbol \cong means "is approximately equal to." Also, from the diagram,

$$MC = MA - CA = s + r$$

and

$$CM' = CA - MA = -r - s' = -(s' + r)$$

Substituting in the above proportion,

$$\frac{s + r}{s} = -\frac{s' + r}{s'}$$

which may easily be put in the form

$$\frac{1}{s} + \frac{1}{s'} = -\frac{2}{r} \quad \text{MIRROR FORMULA} \quad (6b)$$

The primary focal point is defined as that axial object point for which the image is formed at infinity, so substituting $s = f$ and $s' = \infty$ in Eq. 6b we have

$$\frac{1}{f} + \frac{1}{\infty} = -\frac{2}{r}$$

from which

$$\frac{1}{f} = -\frac{2}{r} \quad \text{or} \quad f = -\frac{r}{2} \quad (6c)$$

The secondary focal point is defined as the image point of an infinitely distant object point. That is, $s' = f'$ and $s = \infty$, so that

$$\frac{1}{\infty} + \frac{1}{f'} = -\frac{2}{r}$$

from which

$$\frac{1}{f'} = -\frac{2}{r} \quad \text{or} \quad f' = -\frac{r}{2} \quad (6d)$$

Therefore the primary and secondary focal points fall together, and the magnitude of the focal length is one-half the radius of curvature. When $-r/2$ is replaced by $1/f$, Eq. 6b becomes

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \quad (6e)$$

just as for lenses.

The lateral magnification of the image from a mirror may be evaluated from the geometry of Fig. 6C. From the proportionality of sides in the similar triangles $Q'AM'$ and QAM , we find that $-y'/y = s'/s$, giving

$$m = \frac{y'}{y} = -\frac{s'}{s} \quad (6f)$$

Example: An object 2 cm high is situated 10 cm in front of a concave mirror of radius 16 cm. Find (a) the focal length of the mirror, (b) the position of the image, and (c) the lateral magnification.

Solution: (a) By Eq. 6c,

$$f = -\frac{-16}{2} = 8 \text{ cm}$$

(b) By Eq. 6e,

$$\frac{1}{10} + \frac{1}{s'} = \frac{1}{8} \quad \text{or} \quad \frac{1}{s'} = \frac{1}{8} - \frac{1}{10} = \frac{1}{40}$$

giving

$$s' = 40 \text{ cm}$$

(c) By Eq. 6f,

$$m = -\frac{40}{10} = -4$$

The image occurs 40 cm to the left of the mirror, is four times the size of the object, and is real and inverted.

6.4. Power of Mirrors. The power notation that was used in Sec. 4.9 to describe the image-forming properties of lenses may be readily extended to spherical mirrors as follows. As definitions, we let

$$P = \frac{1}{f}, \quad V = \frac{1}{s}, \quad V' = \frac{1}{s'}, \quad K = \frac{1}{r} \quad (6g)$$

Equations 6b, 6c, 6c, and 6f then take the forms

$$V + V' = -2K \quad (6h)$$

$$V + V' = P \quad (6i)$$

$$P = -2K \quad (6j)$$

$$m = \frac{y'}{y} = -\frac{V}{V'} \quad (6k)$$

Example: An object is located 20 cm in front of a convex mirror of radius 50 cm. Calculate (a) the power of the mirror, (b) the position of the image, and (c) its magnification.

Solution: Expressing all distances in meters, we have

$$K = \frac{1}{0.50} = +2 \text{ D} \quad \text{and} \quad V = \frac{1}{0.20} = +5 \text{ D}$$

By Eq. 6j,

$$P = -2K = -4 \text{ D} \quad \text{Ans. (a)}$$

By Eq. 6i,

$$5 + V' = -4 \quad \text{or} \quad V' = -9 \text{ D}$$

or

$$s' = \frac{1}{V'} = -\frac{1}{9} = -0.111 \text{ m} = -11.1 \text{ cm} \quad \text{Ans. (b)}$$

By Eq. 6k,

$$m = -\frac{5}{-9} = +0.555 \quad \text{Ans. (c)}$$

The power $P = -4 \text{ D}$, and the image is virtual and erect. It is located 11.1 cm to the right of the mirror, and has a magnification of $0.555\times$.

6.5. Thick Mirrors. The term *thick mirror* is applied to a lens system in which one of the spherical surfaces is a reflector. Under these circum-

stances the light passing through the system is reflected by the mirror back through the lens system, from which it emerges finally into the space from which it entered the lens. Of particular importance is a centered system composed of one thin lens and one mirror, separated by an interval d (Fig. 6II). In the same way as for a single mirror, the primary and secondary focal points of such a system coincide at F .

The principal points, which may be found by the oblique-ray construction illustrated in Fig. 6I, coincide also at the position H . In both these diagrams the lens is considered as thin, so that its own principal points may be assumed to coincide at its center. An incident ray parallel to the axis is refracted by the lens, reflected by the mirror, and again refracted by the lens before it crosses the axis of the system at F . The point T where the incident and final rays, when produced, cross each other locates the principal plane and H represents the principal point. If we follow the sign conventions for a single mirror (Sec. 6.3), the focal

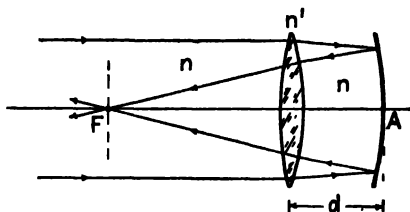


FIG. 6H. A lens and mirror combination is called a "thick mirror."

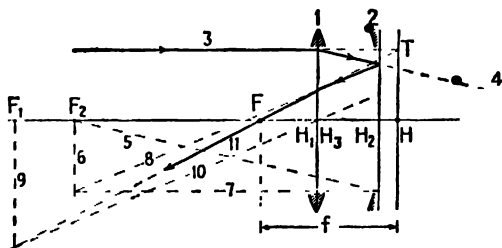


FIG. 6I. Oblique-ray method of construction for a thick-mirror system, locating the focal point F and principal point H .

length f of this particular combination is positive and is given by the interval FH .

6.6. Thick-mirror Formulas. These formulas will be given in the power notation for a case such as that shown in Fig. 6H. Calling r_1 , r_2 , and r_3 the radii of the three surfaces consecutively from left to right, the power of the combination can be shown* to be given by

$$P = (1 - cP_1)(2P_1 + P_2 - cP_1P_2) \quad (6I)$$

* For a derivation of these equations, see J. P. C. Southall, "Mirrors, Prisms, and Lenses," 3d ed., p. 379, The Macmillan Company, New York

where

$$P_1 = (n' - n)(K_1 - K_2) \quad (6m)$$

$$P_2 = -2nK_3 \quad (6n)$$

and

$$K_1 = \frac{1}{r_1}, \quad K_2 = \frac{1}{r_2}, \quad K_3 = \frac{1}{r_3}$$

Of the refractive indices, n' represents that of the lens, and n that of the surrounding space. The distance from the lens to the principal point of the combination is given by

$$H_1H = \frac{c}{1 - cP_1} \quad (6o)$$

where

$$c = \frac{d}{n} \quad (6p)$$

It is important to note from Eq. 6o that the position of H is independent of the power P_2 of the mirror and therefore of its curvature K_3 .

Example: A thick mirror like that shown in Fig. 6H has as one component a thin lens of index $n' = 1.50$ and radii $r_1 = +50$ cm, $r_2 = -50$ cm. This lens is situated 10 cm in front of a mirror of radius -50 cm. Assuming that air surrounds both components, find (a) the power of the combination, (b) the focal length, and (c) the principal point.

Solution: By Eq. 6m the power of the lens is

$$P_1 = (1.50 - 1) \left(\frac{1}{0.50} - \frac{1}{-0.50} \right) = +2 \text{ D}$$

Equation 6n gives for the power of the mirror

$$P_2 = -2 \cdot \frac{1}{-0.50} = +4 \text{ D}$$

From Eq. 6p,

$$c = \frac{d}{n} = \frac{0.10}{1} = 0.10 \text{ m}$$

Finally the power of the combination is given by Eq. 6l as

$$\begin{aligned} P &= (1 - 0.10 \times 2)(2 \times 2 + 4 - 0.10 \times 2 \times 4) \\ &= 0.8(4 + 4 - 0.8) = +5.76 \text{ D} \end{aligned}$$

A power of $+5.76 \text{ D}$ corresponds to a focal length

$$f = \frac{1}{P} = \frac{1}{5.76} = 0.173 \text{ m} = 17.3 \text{ cm}$$

The position of the principal point H is determined from Eq. 6o through the distance

$$H_1H = \frac{0.10}{1 - 0.10 \times 2} = \frac{0.10}{0.80} = 0.125 \text{ m} = 12.5 \text{ cm}$$

It is therefore 12.5 cm to the right of the lens, or 2.5 cm in back of the mirror.

6.7. Special Cases. As a first special case consider a single *thin* lens of which the back surface is silvered. Such a system has the properties of a thick mirror for which the back surface of the lens has the same curvature as the mirror, and the interval d between them is reduced to zero. With these simplifications Eq. 6l reduces to

$$P = 2P_1 + P_2, \quad (6q)$$

and the principal point H coincides with H_1 at the common center of the lens and mirror. P_1 is the power of the thin lens and P_2 is the power of the mirror.

As a second case we may take a *thick* lens silvered on the back. Then Eqs. 6l through 6p may be applied, but for P_1 we use the power of the first surface alone. This is given by Eq. 4k as

$$P_1 = (n' - n)K_1 \quad (6r)$$

P_2 is the power of the second surface as a mirror only, and becomes

$$P_2 = -2n'K_2 \quad (6s)$$

Also

$$c = \frac{d}{n'} \quad (6t)$$

If the lens is considered to be thin, c is set equal to zero and Eqs. 6r and 6s will give the same resultant P as would be obtained by applying Eq. 6q to a lens and mirror.

6.8. Spherical Aberration. The discussion of a single spherical mirror in the preceding sections has been confined to paraxial rays. Within this rather narrow limitation, sharp images of objects at any distance may be formed on a screen, since bundles of parallel rays close to the axis and making only small angles with it are brought to a sharp focus in the focal plane. If, however, the light is not confined to the paraxial region, all rays from one object point do not come to a focus at a common point and we have an undesirable effect known as *spherical aberration*. The

phenomenon is illustrated in Fig. 6J, where parallel incident rays at increasing distances h cross the axis closer to the mirror. The envelope of all rays forms what is known as a *caustic surface*. If a small screen is placed at the paraxial focal plane F and then moved toward the mirror, a point is reached where the size of the circular image spot is a minimum.

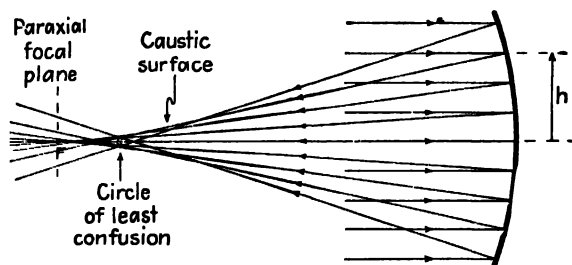


FIG. 6J. Illustrating the spherical aberration of a concave mirror.

This disklike spot is indicated in the diagram and is called the *circle of least confusion*.

The proof that rays from an outer zone of a concave mirror cross the axis inside the paraxial focal point may be simply given by reference to Fig. 6K. According to the law of reflection applied to the ray incident at T , the angle of reflection ϕ' is equal to the angle of incidence ϕ . This in turn is equal to the angle TCA . Having two equal angles, triangle CTX is isosceles, and hence $CX = XT$. Since a straight line is the shortest path between two points,

$$CT < CX + XT$$

Now CT is the radius of the mirror and equals CA , so that

$$CA < 2CX$$

Therefore

$$\frac{1}{2}CA < CX$$

The geometry of the figure shows that as T is moved toward A , the point

FIG. 6K. Geometry showing how marginal rays parallel to the axis of a spherical mirror cross the axis inside the focal point.

X approaches F , and in the limit $CX = XA = FA = \frac{1}{2}CA$.

Over the past years numerous methods of reducing spherical aberration have been devised. If instead of a spherical surface the mirror form is that of a paraboloid of revolution, rays parallel to the axis are all brought to a focus at the same point as in Fig. 6L(a) [see also Sec. 1.6 and Fig. 1E(b)]. Another method is the one shown later in Fig. 10I of inserting

a "corrector plate" in front of a spherical mirror, thereby deviating the rays by the proper amount prior to reflection. With the plate located at the center of curvature of the mirror, a very useful optical arrangement known as the "Schmidt system" is obtained. Still a third system known as a "Mangin mirror," is shown in Fig. 6L(b). Here a meniscus lens is employed in which both surfaces are spherical. When the back surface

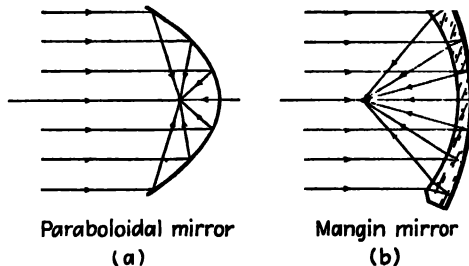


FIG. 6L. Concave mirrors corrected for spherical aberration.

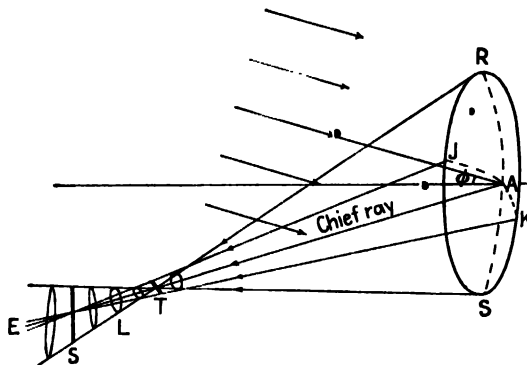


FIG. 6M. Astigmatic images of an off-axis point object at infinity as formed by a concave mirror. The lines T and S are perpendicular to each other.

is silvered to form the concave mirror, all parallel rays are brought to a reasonably good focus.

6.9. Astigmatism. This defect of the image occurs when an object point lies some distance from the axis of a concave or convex mirror. The incident rays, whether parallel or not, make an appreciable angle θ with the mirror axis. The result is that, instead of a point image, two mutually perpendicular line images are formed. This effect is known as astigmatism and is illustrated by a perspective diagram in Fig. 6M. Here the incoming rays are parallel, while the reflected rays are converging toward two lines S and T . The reflected rays in the vertical or *tangential* plane $RASE$ are seen to cross or to focus at T , while the fan

of rays in the horizontal or *sagittal* plane *JAKE* cross or focus at *S*. If a screen is placed at *E* and moved toward the mirror, the image will become a vertical line at *S*, a circular disk at *L*, and a horizontal line at *T*.

If the positions of the *T* and *S* images of distant object points are determined for a wide variety of angles, their loci will form a paraboloidal and a plane surface respectively, as shown in Fig. 6N. As the obliquity of the rays decreases and they approach the axis, the line images not only come closer together as they approach the paraxial focal plane, but they shorten in length. The amount of astigmatism for any pencil of rays is given by the distance between the *T* and *S* surfaces measured along the chief ray.

Equations giving the two astigmatic image positions are as follows:*

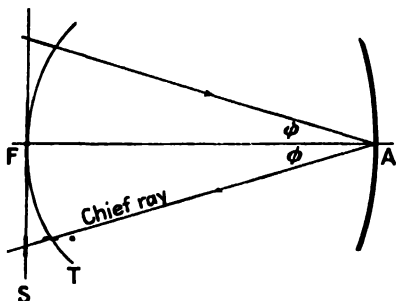


FIG. 6N. Diagram showing the astigmatic surfaces for a concave mirror.

ϕ is the angle of obliquity of the chief ray, and r is the radius of curvature of the mirror.

$$\frac{1}{s} + \frac{1}{s'_r} = \frac{2}{r \cos \phi}$$

$$\frac{1}{s} + \frac{1}{s'_s} = \frac{2 \cos \phi}{r}$$

In both equations s and s' are measured along the chief ray. The angle

The Schmidt optical system, which will be discussed later (Fig. 10I), and the Mangin mirror shown in Fig. 6L(b) constitute instruments in which the astigmatism of a spherical mirror is reduced to a minimum. While the two focal surfaces *T* and *S* exist for these devices, they lie very close together, and the loci of their mean position (such as *L* in Fig. 6M) form a nearly spherical surface. The center of this spherical surface is located at the center of curvature of the mirror as is shown in Fig. 10I.

A paraboloidal mirror, while it is free from spherical aberration even for large apertures, shows unusually large astigmatic *S* — *T* differences off the axis. It is for this reason that paraboloidal reflectors are limited in their use to devices that require a small angular spread, such as astronomical telescopes and searchlights.

Problems

1. The radius of a concave mirror is 50 cm. An object 10 cm high is located in front of the mirror at a distance of (a) 100 cm, (b) 55 cm, (c) 30 cm, and (d) 20 cm. Find the image distance and image size for each of these positions.

* For a derivation of these equations, see G. S. Monk, "Light, Principles and Experiments," 1st ed., pp. 52 and 424, McGraw-Hill Book Company, Inc., New York

2. The radius of a concave mirror is 10 cm. An object 3 cm high is situated in front of the mirror at a distance of (a) 15 cm, (b) 6 cm, (c) 4 cm, and (d) 3 cm. Find the image distance for each of these positions both graphically and by computation. Find also the size of the image in each case.

3. The radius of a convex mirror is 10 cm. An object 3 cm high is situated in front of the mirror at a distance of (a) 10 cm, (b) 6 cm, (c) 5 cm, and (d) 3 cm. Find the image distance for each of these positions both graphically and by computation. Find also the size of the image in each case.

4. A concave mirror is to be used to focus the image of a nearby flower on a wall 9 m from the flower. If a lateral magnification of +10 is desired, what should be the radius of curvature of the mirror?

5. A thin equiconvex lens of index 1.50 and radii 20 cm is silvered on one side. Find the power of this system for light entering the unsilvered side.

6. A thin lens of index 1.50 has as radii $r_1 = +5$ cm and $r_2 = -25$ cm. If the second surface is silvered, what is the power of the system?

7. A thin lens of index 1.50 has as radii $r_1 = -20$ cm and $r_2 = -25$ cm. If the second surface is silvered, what is the power of the system? Use (a) the special-case formula, Eq. 6q, and (b) the thick-lens formulas, Eqs. 6r and 6s, with $d = 0$.

8. A thin lens with a focal length of +10 cm is located 3 cm in front of a concave spherical mirror of radius -20 cm. Find the focal point and principal point of the system, (a) by formula, and (b) graphically.

9. A thin double-concave lens of 20 cm focal length is placed 2 cm in front of a concave mirror of radius -20 cm. Find the focal point and the principal point of the system, (a) by formula, and (b) graphically.

10. A thick lens of index 1.50 has radii $r_1 = +10$ cm and $r_2 = -25$ cm. If the second surface is silvered and the lens is 2 cm thick, find the focal point and the principal point (a) by calculation, and (b) graphically.

11. A lens 1 cm thick, of index 1.60 and radii $r_1 = -4$ cm, $r_2 = -5$ cm, has its second surface silvered as a mirror. Locate the focal point and nodal point (a) by calculation, and (b) graphically.

12. A lens 1 cm thick, of index 1.50 and radii $r_1 = +3$ cm, $r_2 = +10$ cm, has its second surface silvered as a mirror. Locate the focal point and nodal point (a) by calculation, and (b) graphically.

13. An object is located 9 cm in front of a concave mirror of radius 12 cm. Plot a graph of the two astigmatic surfaces from $\phi = 0^\circ$ to $\phi = 30^\circ$.

14. Plot a graph of the two astigmatic surfaces for a concave mirror having a radius of 50 cm. Assume parallel incident light, and show curves from the center out to 30° .

CHAPTER 7

THE EFFECTS OF STOPS

There are two subjects in geometrical optics which, though very important from a practical standpoint, are frequently neglected because they do not directly concern the size, position, and sharpness of the image. One of these is the question of the *field of view*, which determines how much of the surface of a broad object can be seen through an optical system. The other subject is that of the *brightness* of images and the distinction between this, which is important for visual effects, and the illumination, which is important for photographic effects. In treating both the field of view and the brightness of images it is of primary impor-

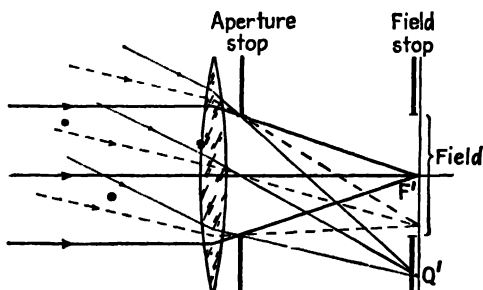


FIG. 7A. Diagram showing the difference between a *field stop* and an *aperture stop*.

tance to understand how and where the bundle of rays traversing the system is limited. The effect of stops or diaphragms, which will always exist if only as the rims of lenses or mirrors, must first be investigated.

7.1. Field Stop and Aperture Stop. In Fig. 7A a single lens with two stops is shown forming the image of a distant object. Three bundles of parallel rays from three different points on the object are shown as brought to a focus in the focal plane of the lens. It may be seen from these bundles that the stop close to the lens limits the size of each bundle of rays, while the stop just in front of the focal plane limits the angle at which the incident bundles can get through to this plane. The first is called an *aperture stop*. It obviously determines the amount of light reaching any given point in the image and therefore controls the brightness of the latter. The second, or *field stop*, determines the extent of the object, or the field, that will be represented in the image.

7.2. Entrance and Exit Pupils. A stop $P'E'L'$ placed behind the lens as in Fig. 7B is in the image space and limits the image rays. By a graphical construction or by the lens formula, the image of this real stop, as formed by the lens, is found to lie at the position PEL shown by the broken lines. Since $P'E'L'$ is inside the focal plane, its image PEL lies in the object space and is virtual and erect. It is called the *entrance pupil*, while the real aperture $P'E'L'$ is, as we have seen, called the *aperture stop*. When it lies in the image space, as it does here, it becomes the *exit pupil*.

It should be emphasized that P and P' , E and E' , and L and L' are pairs of conjugate points. Any ray in the object space directed through one of these points will after refraction pass through its conjugate point

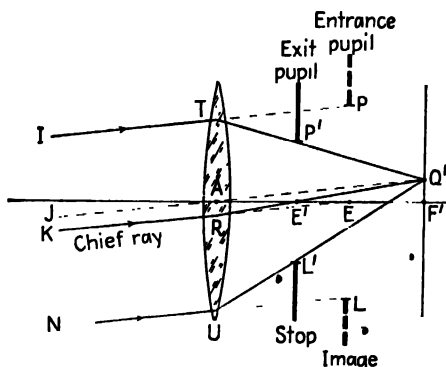


FIG. 7B. Illustrating how an aperture stop and its image become the exit and entrance pupils of a system.

in the image space. Ray IT directed toward P is refracted through P' , ray KR directed toward E is refracted through E' , and ray NU directed toward L is refracted through L' . The image point Q' is located graphically by the broken line JQ' , parallel to the others and passing undeviated through the optical center A . The aperture stop $P'E'L'$ in the position shown also functions to some extent as a field stop, but the edges of the field will not be sharply limited. The diaphragm which acts as a field stop is usually made to coincide with a real or virtual image, so that the edges will appear sharp.

7.3. Chief Ray. Any ray in the object space that passes through the center of the entrance pupil is called a *chief ray*. Such a ray after refraction also passes through the center of the exit pupil. In any actual optical instrument the chief ray rarely passes through the center of any lens itself. The points E and E' at which the chief ray crosses the axis are known as the *entrance pupil point* and the *exit pupil point*. The

former, as we shall see, is particularly important in determining the field of view.

7.4. Front Stop. In certain types of photographic lenses a stop is placed close to the lens, either before it (*front stop*) or behind it (*rear stop*). One of the functions of such a stop, as will be seen in Chap. 9, is to improve the quality of the image formed on the photographic film. With a front stop as shown in Fig. 7C, its small size and its location in the object space make it the entrance pupil. Its image $P'E'L'$ formed by the lens is in the image space and constitutes the exit pupil. Parallel rays IT , JW , and NU have been drawn through the two edges of the entrance pupil and through its center. The lens causes these rays to converge toward the screen as though they had come from the conjugate points P' , E' ,

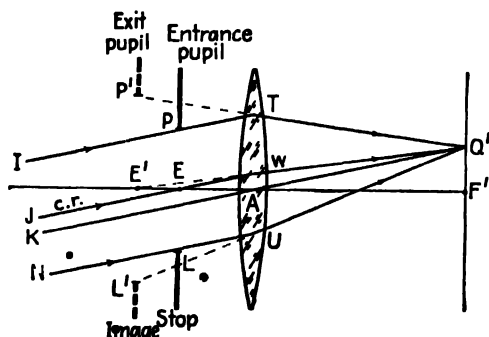


FIG. 7C. A front stop and its image become the entrance and exit pupils of a system.

and L' in the exit pupil. Their intersection at the image point Q' occurs where the undeviated ray KA crosses the secondary focal plane. Note that the chief ray is directed through the center of the entrance pupil in the object space and emerges from the lens as though it had come from the center of the exit pupil in the image space.

While a certain stop of an optical system may limit the rays getting through the system from one object point, it may not be the aperture stop for other object points at different distances away along the axis. For example, in Fig. 7D a lens with a front stop is shown with an object point at M . For this point the periphery of the lens itself becomes the aperture stop, and since it limits the object rays it is the entrance pupil. Its image, which is again the lens periphery, is also the exit pupil. The lens margin is therefore the aperture stop, the entrance pupil, and the exit pupil for the point M . If this object point were to lie to the left of Z , PEL would become the entrance pupil and the aperture stop, and its image $P'E'L'$ the exit pupil.

In the preliminary design of an optical instrument it may not be known which element of the system will constitute the aperture stop. As a result the marginal rays for each element must be investigated one after the other to determine which one actually does the limiting. Regardless of the number of elements the system possesses, it will usually be found to contain but one limiting aperture stop. Once this stop is located, *the entrance pupil of the entire system is the image of the aperture stop formed by all lenses preceding it and the exit pupil is the image formed by*

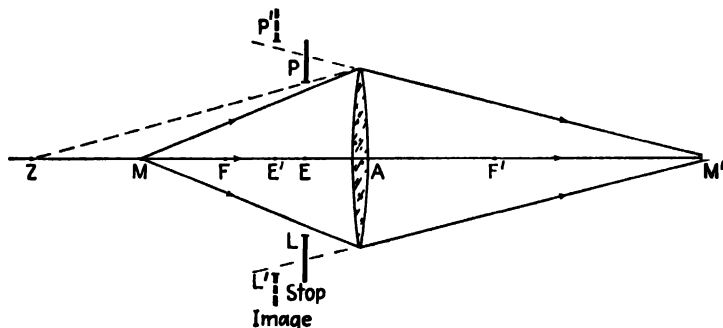


FIG. 7D. The entrance and exit pupils are not the same for all object and image points.

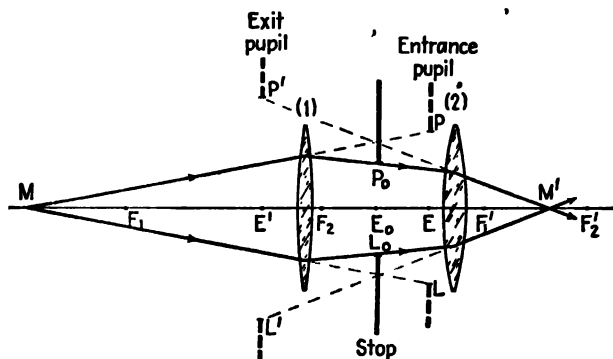


FIG. 7E. Stop between two lenses. The entrance pupil of a system is in the object space of the lens system while the exit pupil is in its image space.

all lenses following it. Figures 7B and 7C, where there is only a single lens either before or behind the stop, should be studied in connection with this statement.

7.5. Stop between Two Lenses. A common arrangement in photographic lenses is to have two separate lens elements with a variable stop or iris diaphragm between them. Figure 7E is a diagram representing such a combination, and in it the elements (1) and (2) are thin lenses while $P_0E_0L_0$ is the stop. By definition the entrance pupil of this system

is the image of the stop formed by lens (1). This image is virtual, erect, and located at PEL . Similarly by definition the exit pupil of the entire system is the image of the stop formed by lens (2). This image, located at $P'E'L'$, is also virtual and erect. The entrance pupil PEL lies in the object space of lens (1), the stop $P_0E_0L_0$ lies in the image space of lens (1) as well as in the object space of lens (2), and the exit pupil $P'E'L'$ lies in the image space of lens (2). Points P_0 and P , E_0 and E , and L_0 and L are conjugate pairs of points for the first lens, while P_0 and P' , E_0 and E' , and L_0 and L' are conjugate pairs for the second lens. This makes points like P and P' conjugate for the whole system. If a point object

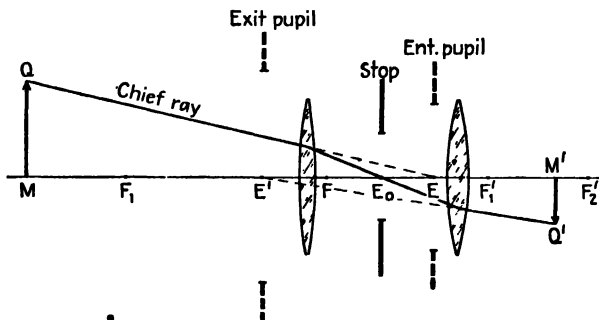


FIG. 7F. The directions of any *chief ray* are such that they pass through the centers of the entrance and exit pupils.

is located on the axis at M , rays MP and ML limit the bundle that will get through the system. At the first lens these rays are refracted through P_0 and L_0 , and at the second lens they are again refracted in such directions that they appear to come from P' and L' as shown. The purpose of using primed and unprimed symbols to designate exit and entrance pupils respectively should now be clear; one lies in the image space, the other in the object space, and they are conjugate images.

The same optical system is shown again in Fig. 7F for the purpose of illustrating the path of a chief ray. Of the many rays that can start from any specified object point Q and traverse the entire system, a chief ray is one which approaches the lens in the direction of E , the entrance pupil point, is refracted through E_0 , and finally emerges traveling toward Q' as though it came from E' , the exit pupil point.

7.6. Two Lenses with No Stop. The theory of stops is applicable not only to cases where circular diaphragms are introduced into an optical system but to any system whatever, since actually the periphery of any lens in the system is a potential stop. In Fig. 7G two lenses (1) and (2) are shown, along with their mutual images as possible stops. Assuming P_1 to be a stop in the object space, its image P' formed by lens (2) lies

in the final image space. Looking upon P_2 as a stop in the image space, its image P formed by lens (1) lies in the first object space. There are therefore two possible entrance pupils, P_1 and P , in the object space of the combination of lenses, and two possible exit pupils, P_2 and P' , in the image space of the combination. For any axial point M lying to the left of Z , P_1 becomes the limiting stop and therefore the entrance pupil of the system. Its image P' becomes the exit pupil. If, on the other hand, M lies to the right of Z , P becomes the entrance pupil and P_2 the exit pupil.

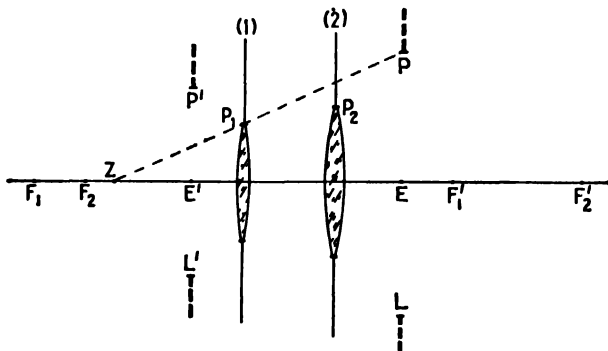


FIG. 7G. The margin of any lens may be the aperture stop of a system.

7.7. Determination of the Aperture Stop. In the system of two lenses with a stop between them represented in Figs. 7E and 7F, the lenses were made sufficiently large so that they did not become aperture stops. If, however, they are not large compared with the stop, as may well be the case with a camera lens when the iris diaphragm is wide open, the system of stops and pupils may become similar to those shown in Fig. 7H. This system consists of two lenses and a stop, each one of which, along with its various images, is a potential aperture stop. P'_1 is the virtual image of the first lens formed by lens (2), P'_0 the virtual image of the stop P formed by lens (2), P_0 the virtual image of P formed by lens (1), and P_2 the virtual image of the second lens formed by lens (1). In other words, when looking through the system from the left, one would see the first lens, the stop, and the second lens in the apparent positions P_1 , P_0 , and P_2 . Looking from the right, one would see them at P'_1 , P'_0 , and P'_2 . Of all these stops P_0 , P_1 , and P_2 are potential entrance pupils located in the object space of the system.

For all axial object points lying to the left of X , P_1 limits the entering bundle of rays to the smallest angle and hence constitutes the entrance pupil of the system. In general the object of which it is the image will be the aperture stop, which in this case is the aperture P_1 of lens (1)

itself. The image of the entrance pupil formed by the entire lens system, namely P'_1 , constitutes the exit pupil. For object points lying between X and Z , P_0 becomes the entrance pupil, P the aperture stop, and P'_0 the exit pupil. Finally, for points to the right of Z , P_2 is the entrance pupil, while P'_2 is both the aperture stop and the exit pupil. It is apparent from this discussion that the aperture stop of any system may change with a change in the object position. The general rule is that *the aperture stop of the system is determined by that stop or image of a stop which subtends the smallest angle as seen from the object point*. If it is determined by an

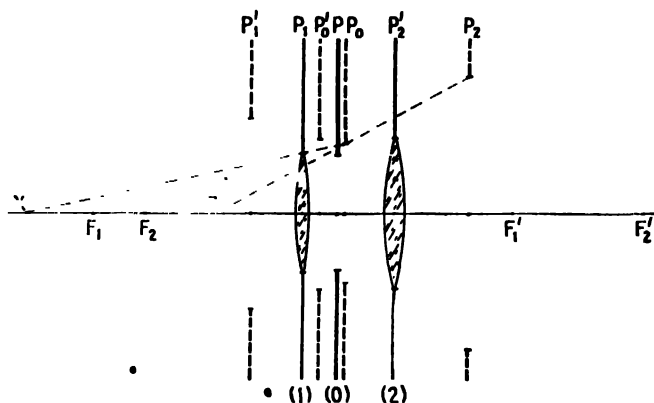


Fig. 7II. A system composed of several parts has many possible stops and pupils.

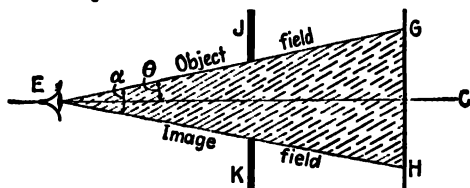


Fig. 7I. Field of view through a window.

image, the aperture stop itself is the corresponding object. In most actual optical instruments the effective stop does not change over the range of object positions normally covered by the instrument when in use.

Having established the methods of determining the positions of the aperture stop and of the entrance and exit pupils, we may now take up the two important properties of an optical system, field of view and brightness. To begin with, let us consider the former property.

7.8. Field of View. When one looks out at a landscape through a window, the field of view outside is limited by the size of the window and by the position of the observer. In Fig. 7I the eye of the observer is shown at E , the window opening at JK , and the observed field at GH .

In this simple illustration the window is the field stop (Sec. 7.1). When the eye is moved closer to the window the angular field α is widened, while when it is moved farther away the field is narrowed. It is common practice with optical instruments to specify the field of view in terms of the angle α and to express this angle in degrees. The angle θ which the extreme rays entering the system make with the axis is called the *half-field angle*, and limits the width of the object that can be seen. This object field includes the angle 2θ , and in this instance is the same as the image field, of angular width α .

7.9. Field of a Plane Mirror. The field of view afforded by a plane mirror is very similar to that of a simple window. As shown in Fig. 7J, TU represents a plane mirror, and $P'E'I'$ the pupil of the observer's eye,

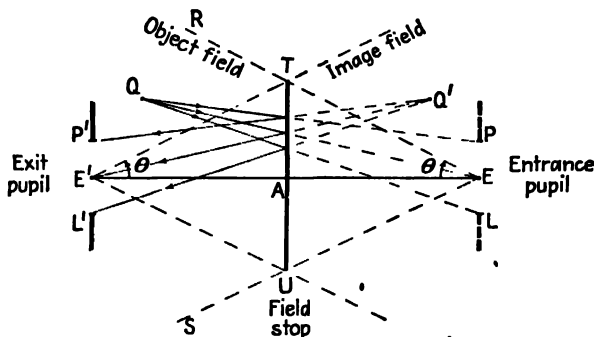


FIG. 7J. Field of view of a plane mirror.

which here constitutes the exit pupil. The entrance pupil PEL is the virtual image of the eye pupil formed by the mirror, and is located just as far behind the mirror as the actual pupil is in front of it. The chief rays $E'T$ and $E'U$ limit the field of view in image space, while the corresponding incident rays ER and ES define the field of view in object space. The latter show the limits of the field in which an object can be situated and still be visible to the eye. As before, it subtends the same angle as does the image field.

The formation of the image of an object point Q within this field is also illustrated. From this point three rays have been drawn toward the points P , E , and L in the entrance pupil. Where these rays encounter the mirror, the reflected rays are drawn toward the conjugate points P' , E' , and I' in the exit pupil. The object Q and the entrance pupil PEL are in the object space, while the image Q' and the exit pupil $P'E'I'$ are in the image space. If Q happens to be located close to RT , only part of the bundle of rays defined by the entrance pupil will be intercepted by the mirror and will be reflected into the exit pupil. In defining the

field of view it is customary to use the chief ray RTE' , although in the present case this distinction is not important because of the relative smallness of the pupil of the eye. Its size is obviously greatly exaggerated in the diagram.

Since the limiting chief ray is directed toward the entrance-pupil point E , the half-field angle θ is in general determined by the smallest angle subtended at E by any stop, or image of a stop, in the system. *The stop determined in this way is the field stop of the system.* For a single mirror the field stop is the border of the mirror itself.

7.10. Field of a Convex Mirror. When the mirror has a curvature the situation is little changed except that the object field and the image field no longer subtend the same angle ($\theta \neq \theta'$ in Fig. 7K). In this figure

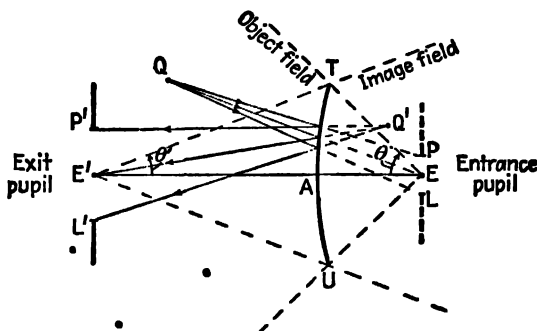


FIG. 7K. Field of view of a convex mirror.

$P'E'L'$ represents the real pupil of an eye placed on the axis of a convex mirror TU . The mirror forms an image PEL of this exit pupil, and this is the entrance pupil which is now smaller in size. Following the same procedure as for a plane mirror, the lines limiting the image field and the object field have been drawn. Rays emanating from an object point Q toward P , E , and L of the entrance pupil are shown as reflected towards P' , E' , and L' in the exit pupil. When extended backward these rays locate the virtual image Q' . The half-field angle θ is here larger than θ' , which determines the field of view to the eye. A similar but somewhat more complicated diagram can be drawn for the field of view of a concave mirror. This case will be left as an exercise for the student, since it is very similar to that of a converging lens to be discussed next.

7.11. Field of a Positive Lens. The method of determining the half-field angles θ and θ' for a single converging lens is shown in Fig. 7L. The pupil of the eye, as an exit pupil, is situated on the right, and its real inverted image appears at the left. The chief rays through the entrance-

pupil point E which are incident at the periphery of the lens are refracted through the conjugate point E' .

The shaded areas, or rather cones, ETU and ERS mark the boundaries within which any object must lie in order to be seen in the image field. The field stop in this case is the lens TU itself, since it determines the half-field angle subtended at the entrance-pupil point. If the eye, and therefore the exit pupil, is moved closer to the lens, thereby increasing the image field angle θ' , the inverted entrance pupil moves to the left, causing a lengthening of the object-field cone ETU .

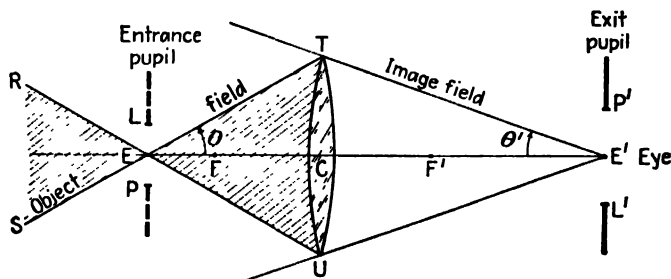


FIG. 7L. Field of view of a converging lens.

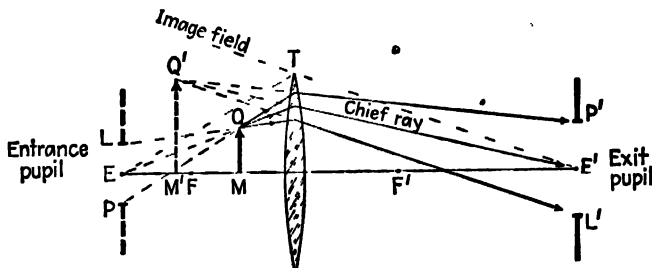


FIG. 7M. Image formation within the field of view of a converging lens.

The same lens has been redrawn in Fig. 7M, where an object QM is shown in a position inside the primary focal point. Through each of the three points P , E , and L , rays are drawn from Q to the lens. From there the refracted rays are directed through the corresponding points P' , E' , and L' on the exit pupil. Extending them backward to their common intersection, the virtual image is located at Q' . The oblique-ray or parallel-ray methods of construction (not shown) may be used to confirm this position of the image. It will be noted that if objects are to be placed near the entrance-pupil point E , they must be very small; otherwise only a part of them will be visible to an eye placed at E' . The student will find it instructive to select object points that lie outside the

object field and to trace graphically the rays from them through the lens. It will be found that invariably they miss the exit pupil.

When a converging lens is used as a magnifier, the eye should be placed close to the lens, since this widens the image-field angle and extends the object field so that the position of the object is less critical.

7.12. Photometric Brightness and Illumination. The amount of light flowing out from a point source Q within the small solid angle subtended by the element of area dA at the distance r [Fig. 7N(a)] is proportional to the solid angle. This is found by dividing the area of dA projected

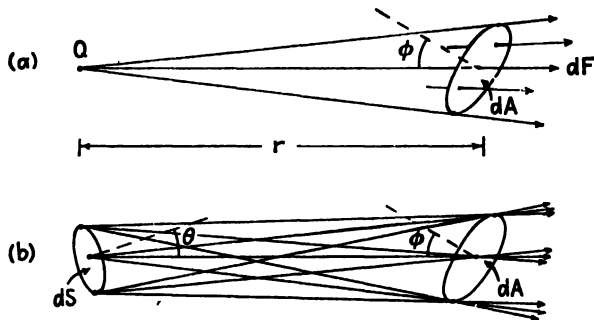


FIG. 7N. An elementary pencil and an elementary beam.

normal to the rays by r^2 , so that the *luminous flux* in this elementary pencil may be written

$$dF = \text{const.} \cdot \frac{dA \cos \phi}{r^2} \quad (7a)$$

Since the source in practice is never a mathematical point, we must consider all pencils emitted from an element of area dS , as shown for three of these pencils in part (b) of Fig. 7N. Assuming that the source is a so-called "Lambert's-law radiator," the flux will now be proportional to the projected area of dS as well, so that

$$dF = \text{const.} \cdot \frac{dS dA \cos \theta \cos \phi}{r^2} \quad (7b)$$

The value of the constant depends only upon the light source, and is called its *photometric brightness* B . To distinguish it from the visual sensation of brightness, it is usually termed the *luminance* in the technical literature, but here we shall use the commoner name brightness, with the understanding that it is the photometric quantity that is meant. The unit of B is experimentally defined as one-sixtieth of the brightness of a black body (Sec. 21.8) at the temperature of melting platinum, and is

called the *candle per square centimeter*. Expressing B in this unit, the flux becomes

$$dF = B \frac{dS \, dA \, \cos \theta \, \cos \phi}{r^2} \text{ lumens} \quad (7c)$$

This is a quantity which must, aside from small losses due to reflection and absorption, remain constant for a bundle of rays as it traverses an optical system.*

The illumination E (also called illuminance) of a surface is defined as the luminous flux incident per unit area, so that

$$dE = \frac{dF}{dA} = \frac{B \cos \theta \, dS \, \cos \phi}{r^2} \quad (7d)$$

Illumination is often expressed in lumens per square meter, or lux. In order to calculate the illumination at any point due to a source having a finite area, we must integrate Eq. 7d over this area:

$$E = \iint \frac{B \, dS \, \cos \theta \, \cos \phi}{r^2} \quad (7e)$$

The exact evaluation of this integral is in general difficult, but in most cases the source is sufficiently far from the illuminated surface so that we may regard both $\cos \phi$ and r^2 as constant. In this case

$$E = \frac{\cos \phi}{r^2} \iint B \cos \theta \, dS = \frac{I \cos \phi}{r^2} \quad (7f)$$

where the integral has been designated by I , since it represents what is called the *luminous intensity* of the source. The definition of this, then, is

$$I = \iint B \cos \theta \, dS \quad (7g)$$

The quantities F , B , E , and I are the four basic ones that are dealt with in the subject of photometry.

As an example appropriate to the present subject let us calculate the illumination due to a luminous disk 6 cm in diameter on a small surface placed normal to the axis of the disk and 20 cm away from it. The brightness of the disk will be taken as 2 candles/cm².

* To be exact, the expression must be multiplied by n^2 in any medium of index n , but since the initial and final media are usually the same, this factor rarely needs to be taken into account.

The condition under which it is legitimate to assume a point source is, by Eq. 7j, that ρ_0^2 shall be negligible with respect to r_0^2 . Even if ρ_0 is as large as $\frac{1}{10} r_0$, the error is only 1 per cent.

7.13. Brightness of an Image. In Fig. 7P is shown a lens forming the image dA' of a surface element dS of the object. If the image is observed by the eye E , the luminous flux dF entering it is limited by the area dA'' of the pupil so that only the narrow bundle indicated by the broken lines contributes to the image on the retina. Now, since the quantity

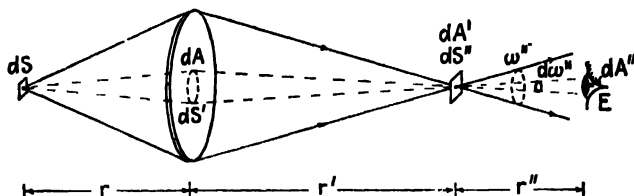


FIG. 7P. Illustrating brightness of an image.

which characterizes a bundle is the multiplier of B in Eq. 7c and since this remains constant through the system, we have, neglecting losses,

$$\begin{aligned} \frac{dF}{B} &= \frac{dS \, dA \cos \theta \cos \phi}{(r)^2} = \frac{dS' \, dA' \cos \theta' \cos \phi'}{(r')^2} \\ &= \frac{dS'' \, dA'' \cos \theta'' \cos \phi''}{(r'')^2} \quad (7k) \end{aligned}$$

The last member of this equation refers to the bundle in the region to the right of the image, and since we assume the flux in the bundle to remain constant and equal to dF , we have

$$\frac{dF}{B} = \frac{dF}{B''} \quad (7l)$$

where B'' denotes the brightness of the image. Hence the important result that

$$B'' = B \quad (7m)$$

For an image formed in the same medium as the object by an optical system in which the losses are negligible, *the brightness of the image equals that of the object.*

This result may seem surprising to one who has experimented at forming images with a lens, because one always finds that when the image is observed on a screen its brightness to the eye increases as the magnification is made smaller. If however the image is observed directly by

the eye, without the use of a screen, its brightness does remain unchanged. This is because the brightness represents the flux per unit area *per unit solid angle*, as can be seen from Eqs. 7k and 7l which give, assuming $\cos \theta'' = \cos \phi'' = 1$,

$$B'' = \frac{dF}{dS'' \frac{dA''}{(r'')^2}} = \frac{dF}{dS'' d\omega''} \quad (7n)$$

When the magnification is decreased the flux incident per unit area of the image is increased, but the total solid angle ω'' (Fig. 7P) is also increased in such a way that the brightness stays constant. The light incident per unit area on a diffusing screen determines its brightness, but this is not the same brightness as the above, since the light is scattered in all directions by such a screen.

7.14. Normal Magnification. In the foregoing discussion, it was assumed that the pupil of the eye acts as the aperture stop of the system. If this is not the case, for example if in Fig. 7P the cone ω'' emerging from the image is not wide enough to fill the eye pupil, the brightness of the image will fall below that of the object. In telescopes and microscopes the eye is usually placed at the exit pupil of the system, and if the full brightness of the object is to be represented in the image, the exit pupil must be at least as large as the pupil of the eye. Now the diameter of the exit pupil is inversely proportional to the magnification, as will be shown for example in the case of a telescope (Eq. 10k). Hence the magnification should not exceed that at which the size of the exit pupil matches that of the eye. This particular value is called the *normal magnification* of the instrument. We shall see that it represents not only the maximum allowable value in order to avoid sacrifice of brightness but also the minimum value required to realize the full resolving power of the instrument (Sec. 15.9).

7.15. Illumination of an Image. The illumination, as defined by Eq. 7e, represents the *total* flux per unit area incident on a surface from all directions. It determines the photographic or other energy effects, as well as the amount of light scattered by unit area of a diffusing screen. To evaluate it in the case of an image formed by a lens or lens system, let us represent this system by A in Fig. 7Q, which also shows the positions of the entrance pupil *PEL* and the exit pupil *P'E'L'*. The brightness B' of the exit pupil as observed at the image point Q' is equal to that of the source, since, from Eq. 7k,

$$\frac{dF}{B} = \frac{dS' dA' \cos \theta' \cos \phi'}{(r')^2} = \frac{dF}{B'}$$

But the brightness is the flux per unit area per unit solid angle, so that if we wish the total flux incident per unit area we must multiply B' by the solid angle ω' subtended by the exit pupil, and this gives

$$E = B'\omega' = B\omega' \quad (7o)$$

Thus the illumination of the image is the product of the brightness of the source and the solid angle subtended by the exit pupil at the image. This relation is not exact, since as may be seen by referring to Eq. 7e, it assumes that all angles are small. It is, however, a good approximation in most actual cases. As in the previous discussion we are here neglecting losses by absorption and reflection. The occurrence of ω' in Eq. 7o is the basis for rating the speed of camera lenses by their f -numbers, as will be explained in Sec. 10.2.

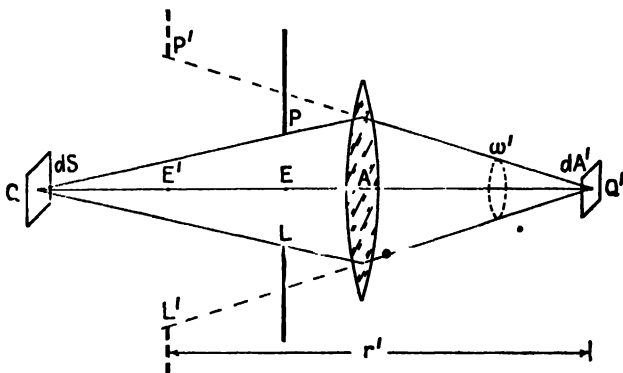


FIG. 7Q. Illustrating the illumination of the image formed by a lens.

It is interesting to note that the illumination is the same as that which would be obtained if the lens were removed and the source were placed at the exit pupil and increased in area to the size of the pupil. The result of the calculation given in Sec. 7.12 may be used to prove this proposition. The illumination due to a disk of brightness B' , the diameter of which subtends a plane angle 2α , was there found to be (see Eq. 7i)

$$E = \pi B' \sin^2 \alpha$$

Provided that α is not too large, a disk of radius $r \sin \alpha$ subtends a solid angle $\omega' = (\pi r^2 \sin^2 \alpha)/r^2 = \pi \sin^2 \alpha$, so that

$$E = B\omega'$$

in agreement with Eq. 7o.

As a practical illustration of this principle, consider the intense beam of light produced by a spotlight or searchlight, as illustrated in Fig. 7R. The rim of the reflector of aperture A is the entrance pupil as well as

the exit pupil. Neglecting losses of light by reflection and absorption, the illumination over the region D on a distant screen M is the same as that which would be obtained were the reflector removed and a source of the same brightness as S but having the full size of A placed at the position of A . The equivalent *beam candle power* of a spotlight or searchlight is defined as the candle power of a bare source which, if

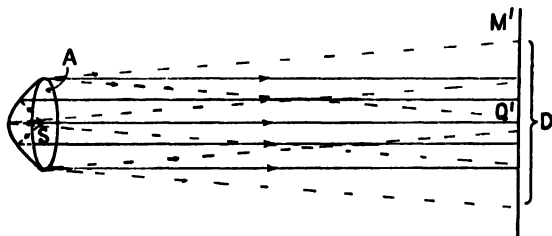


FIG. 7R. A spotlight or searchlight beam is often rated in terms of its beam candle power. located at the same distance away from a given point, would produce at that point the same illumination.

7.16. Image of a Point Source. The above principle is applicable to the illumination in the image of a source of finite area. If the area of the source is negligible, as it is for example in the telescopic images of stars, the principle deduced above is no longer applicable. The image,

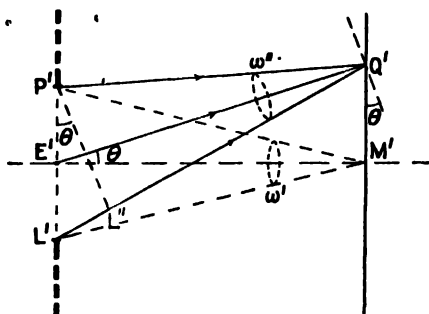


FIG. 7S. Illumination at an off-axis point in the image.

instead of being of the very small size predicted by geometrical optics, is actually broadened because of diffraction by the aperture of the lens system (Sec. 1.1). Its illumination is therefore less than would be predicted by Eq. 7o. The investigation of this case requires the results of the theory of diffraction and will therefore be postponed until we take up this subject (Sec. 15.10).

7.17. Illumination off the Axis. Supposing the object were a plane surface of uniform brightness, it would be found that the illumination in the image would fall off with distance away from the axis. This effect is due to more than one cause. In Fig. 7S let $P'E'L'$ represent the exit pupil, which has a uniform brightness B' equal to that of the source. At the axial point M' the illumination is, according to Eq. 7o, equal to $B'\omega'$. For a point such as Q' , however, the following factors act to decrease the illumination: (a) a factor $\omega''/\omega' = \cos^2 \theta$; (b) a factor $P'L''/P'L' = \cos \theta$, representing the decrease in area of the exit pupil as

seen from Q' compared to that seen from M' ; and (c) another factor $\cos \theta$ coming from the fact that the light is not incident normally on the surface at Q' , as it would be on the surface represented by the broken line. Tilting a surface through an angle θ distributes the flux over an area which is $1/\cos \theta$ times larger, and hence the illumination, or flux per unit area, is decreased by $\cos \theta$. Putting all these factors together, we have, for the illumination at Q' ,

$$E'' = B'\omega' \cos^4 \theta \quad (7p)$$

Near the axis the factor $\cos^4 \theta$ varies only slightly from unity, but if α becomes as great as 30° , for example, the illumination is reduced by 44 per cent.

7.18. Vignetting. Another effect, which may cause the illumination off the axis to fall at an even more rapid rate, is that known as *vignetting*.

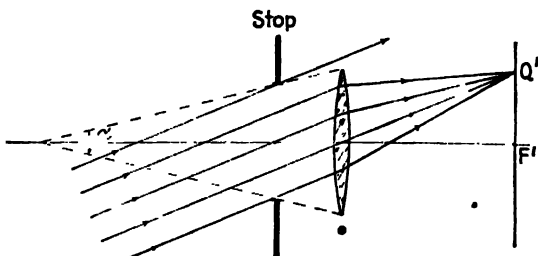


FIG. 77. Illustrating the property called *vignetting*.

This is particularly likely to occur in a lens system containing stops, as is illustrated for a single lens in Fig. 77'. Although the aperture of the stop is smaller than that of the lens, at the angle of incidence shown some of the rays at the top miss the lens entirely, while the lower part of the lens receives no light. For distant object points, the field that is reproduced without vignetting covers angles up to the value of α shown in the diagram. At wider angles the field begins to darken more rapidly than would be indicated by Eq. 7p. Vignetting is seldom encountered in telescopes or in other instruments having a relatively small field of view, but in instruments like wide-angle cameras it can become serious.

Problems

1. A thin double-convex lens of focal length 5 cm and aperture 5 cm has a 2-cm stop located 2 cm in front of it. An object 2 cm high is situated with its lower end on the lens axis 10 cm in front of the lens. (a) Locate graphically and by formula the position and size of the exit pupil. (b) Locate the image of the object graphically by drawing the two marginal rays and the chief ray from the top end of the object.
2. A thin double-convex lens of focal length 6 cm and aperture 5 cm has a 2-cm stop located 2.5 cm behind it. An object 4 cm high is situated with its lower end on the lens axis 14 cm in front of the lens. (a) Find graphically and by formula the

position and size of the entrance pupil. (b) Locate the image graphically by drawing the two marginal rays and the chief ray from the top end of the object.

3. A thin double-concave lens of focal length 6 cm and aperture 5 cm has a 2 cm stop located 4 cm in front of it. An object 2 cm high is situated with its lower end on the lens axis 14 cm in front of the lens. (a) Find graphically the two pupils. (b) Graphically locate the image and draw the two marginal rays and the chief ray from the top end of the object.

4. Two thin lenses with focal lengths of +10 cm and +5 cm respectively and with apertures of 4 cm are spaced 4 cm apart. A stop 2 cm in diameter is located midway between the lenses, and an object 4 cm high is located 10 cm in front of the first lens. (a) Find by formula and graphically the entrance and exit pupils. (b) Locate the final image and show the two marginal rays and the chief ray from the top end of the object.

5. A thin lens of focal length +8 cm and aperture 4 cm is located 3 cm in front of another thin lens of focal length +6 cm and aperture 3 cm. (a) Locate graphically and label all the pupils of the system. (b) An object 1 cm high is located 6 cm in front of the first lens. Find the final image and show the two marginal rays and the chief ray from the top of the object. Label the entrance pupil and the exit pupil.

6. An exit pupil with a 3-cm aperture is located 6 cm in front of a convex spherical mirror of 10 cm radius. An object 3 cm high is placed with its center on the axis 4 cm in front of the mirror. Find graphically (a) the entrance pupil, (b) the image, and (c) the minimum aperture required for the mirror in order to be able to see the entire object from all points of the exit pupil.

7. An exit pupil with a 3-cm aperture is located 15 cm in front of a concave spherical mirror of 10 cm radius. An object 1 cm high is centrally located on the axis 12 cm in front of the mirror. Find graphically (a) the entrance pupil, (b) the image, and (c) the minimum aperture for the mirror required in order to see the entire object.

8. Construct to scale a diagram of the field of view for a positive lens used as a magnifier. Assume a focal length $f = 5$ cm, an exit pupil of width 1 cm situated 3 cm from the lens, and an object 2 cm high centrally located 3 cm from the lens on the opposite side.

9. Calculate the illumination in lux due to a frosted lamp bulb of projected area 30 cm² and average brightness 0.955 candle/cm² on a surface normal to the light and 10 m away. (NOTE: Because of Lambert's law the bulb may be treated as a flat surface of the area and brightness given.)

10. If the lamp in Prob. 9 is displaced 5 m in a direction at right angles to the original line joining it and the illuminated surface, what will be the new value of the illumination?

11. A convex lens of focal length 20 cm and aperture 5 cm has a stop 2 cm in diameter situated 3 cm in front of it. A small disk of brightness 6 candles/cm² is placed centrally on the axis 60 cm from the lens. Calculate (a) the illumination at the image, (b) the size of the exit pupil, and (c) the angle at which vignetting begins.

12. A wide-angle camera has a lens of focal length 10 cm, and takes photographs on a 9-by-12 cm film. Assuming no vignetting, find the fraction by which the exposure is diminished at the corners of the film.

13. The diameter of a thin lens is 2 cm and it has a focal length of 20 cm. If this lens is placed halfway between the eye and a large object 18 cm from the eye, find what width of the object can be seen through the lens.

14. If the object in Prob. 2 has a brightness of 3 candles/cm², what is the illumination at the image?

CHAPTER 8

RAY TRACING

The discussion of image formation by a system of one or more spherical surfaces has up to this point been confined to the consideration of paraxial rays. With this limitation it has been possible to derive relatively simple methods of calculating and constructing the position and size of the image. In practice the apertures of most lenses are so large that paraxial rays constitute only a very small fraction of all the effective rays. It is therefore important to consider what happens to rays that are not paraxial. The straightforward method of attacking this problem is to trace the paths of the rays through the system, applying Snell's law to the refraction at each surface. The fundamental equations for this procedure will be presented in this chapter, and a sample calculation will be carried through for a single thick lens.

8.1. Oblique Rays. All rays which are not paraxial are called *oblique rays*. When the law of refraction is accurately applied to the paths of all rays from one object point as they pass through one or more coaxial surfaces, the position of the image point is found to vary somewhat with the obliquity of the rays. This leads to a blurring of the image known as a *lens aberration*, and the study of these aberrations will be the subject of the following chapter. Experience shows that it is possible, by properly choosing the radii and positions of spherical refracting surfaces, to reduce the aberrations greatly. Only in this way can optical instruments be designed having large usable apertures and at the same time good image-forming qualities.

Lens designers follow two general lines of approach to the problem of finding the optimum conditions. The first is to use well-known aberration formulas to calculate the approximate radii and spacing of the surfaces that are desirable for the particular problem. The second and most rigorous test is to apply the exacting methods of calculation known as *ray tracing* to find the paths of several representative rays. Some of these rays will be paraxial and some oblique, and each is traced through to the image. If the results are not satisfactory, the surfaces are moved, the radii are changed, and the process is repeated until an apparent minimum of aberration is obtained. This is a long and tedious cut-and-try process, requiring in some cases hundreds of hours of work. Five-

six-, or even seven-place logarithms may be required, and certain standard tabular forms are printed by the different designers for recording the calculations and results (see Table 8I). We shall first consider the derivation of the equations needed in this process.

8.2. Ray-tracing Formulas. A diagram from which these formulas may be derived is given in Fig. 8A. An oblique ray MT making an angle θ with the axis is refracted by the single spherical surface at T so that it crosses the axis again at M' . The line TC is the radius of the

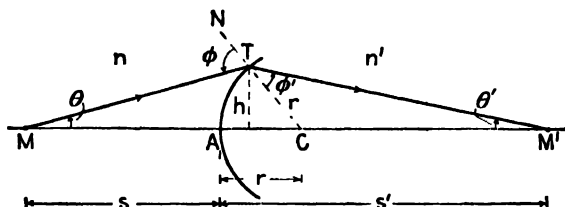


FIG. 8A. Refraction of a ray at a spherical surface, showing the geometry used in ray tracing.

refracting surface and constitutes the normal from which the angles of incidence and refraction at T are measured. As regards the signs of the angles involved, we consider that

1. Slope angles are positive when the axis must be rotated counter-clockwise through an angle of less than $\pi/2$ to bring it into coincidence with the ray.
2. Angles of incidence and refraction are positive when the radius of the surface must be rotated counterclockwise through an angle of less than $\pi/2$ to bring it into coincidence with the ray.

Accordingly, angles θ , ϕ , and ϕ' in Fig. 8A are positive, while angle θ' is negative.

Applying the law of sines to the triangle MTC , one obtains

$$\frac{\sin(\pi - \phi)}{r + s} = \frac{\sin \theta}{r}$$

Since the sine of the supplement of an angle equals the sine of the angle itself,

$$\frac{\sin \phi}{r + s} = \frac{\sin \theta}{r}$$

Solving for $\sin \phi$, we find

$$\sin \phi = \frac{r + s}{r} \sin \theta \quad (8a)$$

Now by Snell's law the angle of refraction ϕ' in terms of the angle of incidence ϕ is given by

$$\sin \phi' = \frac{n}{n'} \sin \phi \quad (8b)$$

In the triangle MTM' the sum of all interior angles must equal π . Therefore

$$\theta + (\pi - \phi) + \phi' + (-\theta') = \pi$$

which, upon solving for θ' , gives

$$\theta' = \phi' + \theta - \phi \quad (8c)$$

This equation allows us to calculate the slope angle of the refracted ray. To find where the ray crosses the axis and the image distance s' , the law of sines may be applied to the triangle TCM' , giving

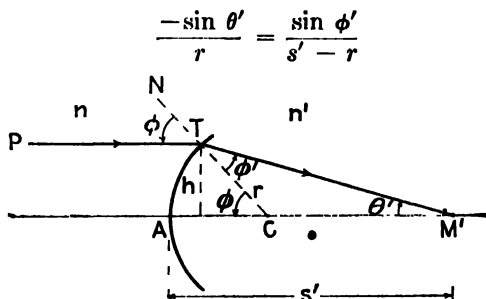


Fig. 8B. Showing the geometry of ray tracing for incident parallel light.

The image distance is therefore

$$s' = r - r \frac{\sin \phi'}{\sin \theta'} \quad (8d)$$

An important special case is that in which the incident ray is parallel to the axis. Under this simplifying condition it may be seen from Fig. 8B that

$$\sin \phi = \frac{h}{r} \quad (8e)$$

where h is the height of the incident ray PT above the axis. For the triangle TCM' , the sum of the two interior angles ϕ' and θ' equals the exterior angle at C . When the angles are assigned their proper signs, this gives

$$\theta' = \phi' - \phi \quad (8f)$$

The six of the above equations which are numbered form an important set by which any ray lying in a *meridian plane* may be traced through

a number of coaxial spherical surfaces. A meridian plane is defined as any plane containing the axis of the system. While most of the rays emanating from an extraaxial object point do not lie in a meridian plane, the image-forming properties of an optical system can usually be determined from properly chosen meridian rays. *Skew rays*, or rays that are not confined to a meridian plane, do not intersect the axis and are difficult to trace.

8.3. Sample Ray-tracing Calculations. These will be illustrated in the case of an equiconvex lens with radii $r_1 = +10$ cm and $r_2 = -10$ cm, the lens being made of crown glass having an index $n' = 1.52300$ for the Fraunhofer D line. If the axial thickness is 2 cm, let us find the focal points for parallel rays incident at heights above the axis $h = 1.5$ cm, 1.0 cm, and 0 cm.

A diagram for this problem is given in Fig. 8C. Refraction at the first surface directs the ray toward the corresponding image point at

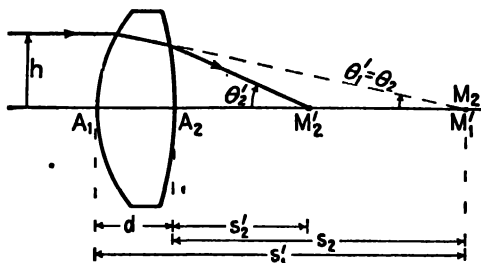


FIG. 8C. Path of a parallel ray through a double-convex lens.

M_2' . This becomes the object point M_2 for the second surface, at which the refraction determines the final image point M_1' . The following two tables give the calculations for the refracting surfaces separately. For the first surface the incident light is parallel to the axis, and the four ray-tracing formulas to be used are Eqs. 8b, 8d, 8e, and 8f, namely

$$\sin \phi = \frac{h}{r}, \quad \sin \phi' = \frac{n}{n'} \sin \phi,$$

and

$$c' = \phi' - \phi, \quad r - s' = r \frac{\sin \phi'}{\sin \theta'}.$$

By substitution of the known values of h and r_1 in the first equation, $\sin \phi$ is determined. Inserting this, along with the known n and n' , in the second equation, we obtain $\sin \phi'$. Having found ϕ and ϕ' , we may use the third equation to calculate θ' . Finally the values of r_1 , ϕ' , and θ' are used in the last equation to obtain the image distance s_1' . In the

use of logarithms for these calculations, the subtraction of one logarithm from another to find a quotient is avoided by employing the cologarithms of all quantities occurring in the denominator. Thus the operations are reduced to those of addition. The procedure is self-evident as regards the first two columns in Table 8I, where it is shown for the first refracting surface.

TABLE 8I. • CALCULATIONS FOR THE FIRST SURFACE
 $r = +10.0$ cm, $n' = 1.52300$, $n = 1.00000$

	$h = 1.5$ cm	$h = 1.0$ cm	$h = 0$
$\log h$	0.176091	0.000000	
$\text{colog } r$	9.000000	9.000000	
$\log \sin \phi$	9.176091	9.000000	
$\log n$	0.000000	0.000000	
$\text{colog } n'$	9.817300	9.817300	
$\log \sin \phi'$	8.993391	8.817300	
ϕ'	5°39'6"	3°45'53"	0.098490
ϕ	8°37'37"	5°44'21"	0.150000
$\theta' (= \phi' - \phi)$	2°58'31"	1°58'28"	0.051510
$\text{colog } \sin \theta'$	1.284790	1.462767	1.288108.
$\log \sin \phi'$	8.993391	8.817300	8.993391
$\log r$	1.000000	1.000000	1.000000
$\log (r - s')$	1.278181	1.280067	1.281499
$r - s'$	-18.9750	-19.0576	-19.1205
s'	+28.9750 cm	+29.0576 cm	+29.1205 cm

For the case $h = 0$ in the right-hand column a special procedure is followed. The calculation is started by first finding the number that corresponds to one of the values of $\log \sin \phi'$ in another column. In the present case either column may be used. This number is entered opposite ϕ' in the table under $h = 0$. For example, in the column headed $h = 1.5$ cm, we find $\log \sin \phi' = 8.993391$, and the number corresponding to this logarithm (namely, 0.098490) is entered in the last column. Following the same procedure for the corresponding angle ϕ we find $\log \sin \phi = 9.176091$, and the number 0.150000 is shown opposite ϕ . The difference between these two numbers is next entered for θ' . Then opposite $\text{colog } \sin \theta'$ is written the cologarithm of the number 0.051510,

namely 1.288108. From this point on the procedure is the same as for the auxiliary ray chosen, values of $\log \sin \phi'$ and $\log r$ being taken from the column originally selected. The value of s' that results will be the same whatever auxiliary ray is chosen for the calculation.*

It is to be noted that the image distance s' is greatest for $h = 0$, and about one-half of 1 per cent less for the 1.5-cm ray. These slightly different image points M'_1 become the object points for the second lens surface, and the slope angles θ' of Table 8I become the slope angles θ for the incident rays in Table 8II. Since in the latter case the object rays are not parallel to the axis, the four ray-tracing formulas to be used for Table 8II are Eqs. 8a, 8b, 8c, and 8d, namely

$$\sin \phi = \frac{r}{r} + \frac{s}{r} \sin \theta, \quad \sin \phi' = \frac{n}{n'} \sin \phi,$$

and

$$\theta' = \phi' + \theta - \phi, \quad r - s' = r \frac{\sin \phi'}{\sin \theta'}$$

Starting with the first equation, r_2 is given as -10.0 cm and s is the distance A_2M_2 in Fig. 8C. It is obtained by subtracting from the values of s' in Table 8I the lens thickness $d = 2.0$ cm.

Taking the ray having $h = 1.5$ cm as an example, we have s' from Table 8I as 28.9750 cm, which after subtraction of $d = 2.0$ cm yields $s = -26.9750$ cm. The negative sign signifies that the object ray corresponds to a virtual object. Since both r_2 and s have negative signs, the two magnitudes are added in the first equation to give -36.9750 cm. In the last column of Table 8II the value for $\log \sin \theta = 8.711892$ is obtained as $\text{colog} \sin \theta' = 1.288108$ of Table 8I. Instead of angles ϕ' , θ , and ϕ in the last column, numbers are obtained by the auxiliary-ray method described above in connection with Table 8I. For example, the number 0.291210 corresponds to $\log \sin \phi' = 9.464206$ and is entered opposite ϕ' in Table 8II. The number 0.051510 corresponds to $\log \sin \theta = 8.711892$ and is entered opposite θ in the table. Then opposite $\text{colog} \sin \theta'$ is written 0.819550, which is the cologarithm of the number 0.151513. From here on, the procedure corresponds to that for the other rays.

The final figures show that when parallel rays are incident on the lens of Fig. 8C at heights of 1.5 cm, 1.0 cm, and 0 cm, the axial intercepts are at 8.879 cm, 9.088 cm, and 9.220 cm, respectively. Thus the distance

* The theory of this auxiliary ray method is set forth in Lummer, "Photographic Optics," English translation by S. P. Thompson, p. 126, The Macmillan Company, New York

TABLE 8II. CALCULATIONS FOR THE SECOND SURFACE
 $r = -10.0$ cm, $n' = 1.00000$, $n = 1.52300$

	$\theta = 2^{\circ}58'31''$	$\theta = 1^{\circ}58'28''$	$\theta = 0$
$r + s$	-36.9750	-37.0576	-37.1205
$\log (r + s)$	1.567908	1.568877	1.569614
$\text{colog } r$	9.000000	9.000000	9.000000
$\log \sin \theta$	8.715127	8.537233	8.711892
$\log \sin \phi$	9.283035	9.106110	9.281506
$\log n$	0.182700	0.182700	0.182700
$\text{colog } n'$	0.000000	0.000000	0.000000
$\log \sin \phi'$	9.465735	9.288810	9.464206
ϕ'	16°59'31''	11°12'32''	0.291210
θ	2°58'31''	1°58'28''	0.051510
ϕ	11°3'45''	7°20'7''	0.191207
$\theta' (= \phi' + \theta - \phi)$	8°54'17''	5°50'53''	0.151513
$\text{colog } \sin \theta'$	0.810252	0.991955	0.819550
$\log \sin \phi'$	9.465735	9.288810	9.464206
$\log r$	1.000000	1.000000	1.000000
$\log (r - s')$	1.275987	1.280765	1.283756
$r - s'$	-18.879	-19.088	-19.220
s'	+8.879cm	+9.088cm	+9.220cm
$\delta s'$	0.341	0.132	0

from the lens vertex to the second focal point is not a constant but varies slightly for different zones of the lens. This defect is called *spherical aberration* and will be discussed in detail in the next chapter. The focal distances s' for $h = 0$ and for $\theta = 0$ in Tables 8I and 8II are identical with the values which would be obtained from the paraxial ray formulas given in Sec. 4.8.

8.4. First-order Theory. In four of the six ray-tracing formulas of the preceding section, the sines of the angles appear rather than the angles themselves. Expansion of the sine of an angle by Maclaurin's theorem gives

$$\sin \theta = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \quad (8g)$$

For paraxial rays the slope angles are very small and we may, to a first approximation, neglect all terms but the first and write $\sin \theta = \theta$. (For an angle of 5.73° , or $\frac{1}{10}$ rad, the first term $\theta = 0.100000$, while the second term $\theta^3/3! = 0.000167$.) The theory of lenses based upon the neglect of all but the first term in this expansion is an approximate theory known as *first-order theory* or *Gauss theory*.

When θ is small, the other angles in Eqs. 8a, 8b, 8c, and 8d (namely, θ' , ϕ , and ϕ') are also small, provided the ray lies close to the axis, as it must to be paraxial. The ray-tracing formulas for this type of ray therefore reduce to

$$\begin{aligned} \phi &= \frac{r+s}{r} \theta, & \phi' &= \frac{n}{n'} \phi, \\ \theta' &= \phi' + \theta - \phi, & \text{and } s' &= r - r \frac{\phi'}{\theta} \end{aligned} \quad (8h)$$

By algebraic substitution of the first formula in the second, of the resultant formula in the third, and of this resultant in the fourth, all angles may be eliminated. The final equation can then be simplified to give the Gaussian formula, Eq. 4a:

$$\frac{n}{s} + \frac{n'}{s'} = \frac{n' - n}{r} \quad (8i)$$

which therefore corresponds to first-order theory.

8.5. Abbe's Sine Condition. In Chap. 4 it was found that the lateral magnification produced by a single spherical surface was given by the relation (Eq. 4f):

$$m = \frac{y'}{y} = - \frac{s' - r}{s + r} \quad (8j)$$

This equation, which follows from the similarity of triangles MQC and $M'Q'C$ in Fig. 8D, is clearly valid whether or not the ray QCQ' is paraxial. From Eq. 8a we obtain

$$s + r = r \frac{\sin \phi}{\sin \theta}$$

and from Eq. 8d,

$$s' - r = -r \frac{\sin \phi'}{\sin \theta'}$$

Substitution of these quantities in the right-hand side of Eq. 8j yields

$$\frac{y'}{y} = \frac{\sin \phi'}{\sin \theta'} \frac{\sin \theta}{\sin \phi} \quad (8k)$$

Now, according to Snell's law,

$$\frac{\sin \phi'}{\sin \phi} = \frac{n}{n'}$$

which upon substitution in Eq. 8k gives

$$\frac{y'}{y} = \frac{n}{n'} \frac{\sin \theta}{\sin \theta'}$$

or

$$ny \sin \theta = n'y' \sin \theta' \quad \text{SINE CONDITION} \quad (8l)$$

Since nowhere in its derivation has the assumption of paraxial rays been introduced, this relation holds for arbitrarily large values of θ and θ' . It

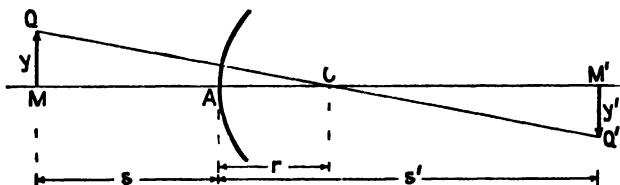


FIG. 8D. The undeviated ray through a single spherical surface.

is one of the most important equations of geometrical optics, as will appear in the following chapter.

The sine condition written for paraxial rays becomes simply

$$ny \theta = n'y' \theta' \quad (8m)$$

a relation often referred to as *Lagrange's law*. In both Eqs. 8l and 8m all quantities on the left side refer to the object space, while all those on the right side refer to the image space. Also, if we restrict θ and θ' to small angles, Fig. 8A shows that they may be written

$$\theta = \frac{h}{s} \quad \text{and} \quad \theta' = \frac{h}{s'}$$

Substitution of these values in Eq. 8m gives for the magnification

$$m = \frac{y'}{y} = \frac{n\theta}{n'\theta'} = -\frac{ns'}{n's} \quad (8n)$$

Since Abbe's sine condition applies separately to each spherical refracting surface, it may be applied in succession to a number of surfaces. Hence one concludes that $ny \sin \theta$ may be taken as applying to the object space of a complete optical system and $n'y' \sin \theta'$ to the final image space of that system.

Problems

1. A double-convex lens is made of borosilicate crown glass of index 1.50000. It has radii $r_1 = +10$ cm and $r_2 = -10$ cm and is 1 cm thick. Using five-place logarithms, locate the focal points for (a) $h = 1.5$ cm, (b) $h = 1.0$ cm, (c) $h = 0.5$ cm, and (d) $h = 0$.
2. Solve Prob. 1 when $r_1 = +6.25$ cm and $r_2 = -25$ cm.
3. Solve Prob. 1 when $r_1 = +5$ cm and $r_2 = +\infty$.
4. Solve Prob. 1 when $r_1 = +3.75$ cm and $r_2 = +15$ cm.
5. Solve Prob. 1 when $r_1 = +25$ cm and $r_2 = -6.25$ cm.
6. Solve Prob. 1 when $r_1 = \infty$ and $r_2 = -5$ cm.
7. Solve Prob. 1 when $r_1 = -15$ cm and $r_2 = -3.75$ cm.
8. From the paraxial ray-tracing formulas, Eqs. 8h, derive the Gaussian formula for a single surface, Eq. 8i.
9. Solve Prob. 1 for an object point located on the axis 20 cm in front of the first vertex of the lens.

CHAPTER 9

LENS ABERRATIONS

The description of ray tracing that was given in the last chapter served to emphasize the limitations of Gauss' first-order theory. A wide beam of rays striking a lens parallel to its axis was found not to be focused at a unique point. The first-order theory upon which the simple formulas of the preceding chapters were based will in general give only an idealized picture of the actual conditions encountered with lenses of wide aperture and wide field of view. When ray tracing is applied to points off the axis, the phenomenon of spherical aberration gives way to no less than four other defects of the image. The reduction of these aberrations to a minimum is one of the chief problems of geometrical optics and also one of the most complicated. It would be impossible to set forth within the scope of a single chapter any details of the extensive theory involved in this problem.* Instead we shall describe how each aberration manifests itself and present some of the results of the theory as to how it may be diminished.

9.1. The Five Seidel Aberrations. Since aberrations represent departures from the simple first-order theory, the natural way to investigate them is to take account not only of the first term in the expansion of the sine (Eq. 8g) but also of the second term and even of the third. Since the second term contains the angle raised to the third power, inclusion of it leads to the so-called *third-order theory*. This involves replacing the sines of the angles in the ray-tracing formulas by the first two terms of the expansion. Thus $\sin \theta$ becomes $\theta - (\theta^3/3!)$, $\sin \phi$ becomes $\phi - (\phi^3/3!)$, etc. The resulting equations give a reasonably accurate account of the principal aberrations.

In this theory the aberration for any ray, *i.e.*, its deviation from the path prescribed by the Gauss theory, is expressed in terms of five sums, S_1 to S_5 , called the *Seidel sums*. If a lens were to be free of all aberration, all five of these would have to be simultaneously and individually equal to zero. No optical system can be made to satisfy all these conditions at once. Therefore it is customary to treat each sum separately, and

* For a thorough account of lens aberrations, the reader is referred to A. E. Conrady, "Applied Optics and Optical Design," Vol. I, Oxford University Press, New York.

the vanishing of certain ones corresponds to the absence of certain aberrations. Thus, if for a given axial object point the first Seidel sum $S_1 = 0$, there is no *spherical aberration* at the corresponding image point. If both $S_1 = 0$ and $S_2 = 0$, the system will also be free of *coma* and Abbe's sine condition (Eq. 81) will be satisfied. If in addition to $S_1 = 0$ and $S_2 = 0$, the sums $S_3 = 0$ and $S_4 = 0$ as well, the images will be free of *astigmatism* and *curvature of field*. If finally S_5 could be made to vanish, there would be no *distortion* of the image. These aberrations are also known as the *five monochromatic aberrations* because they exist for any specified color and refractive index. Additional image defects occur when the light

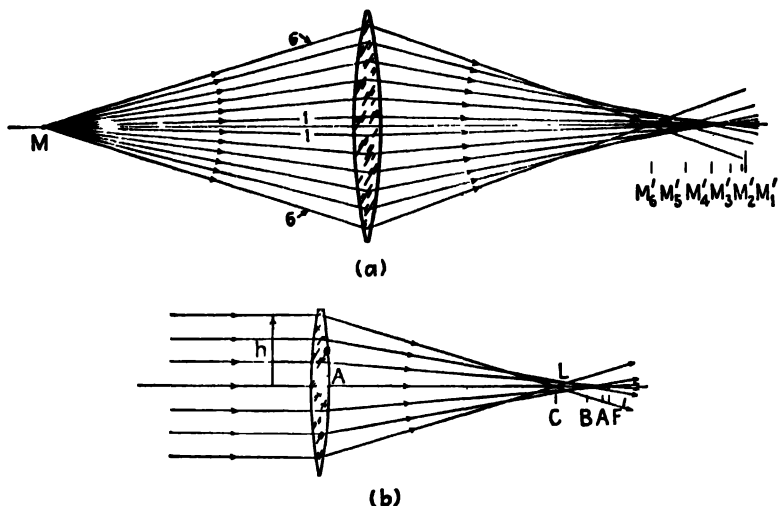


FIG. 9A. Illustrating the spherical aberration of a single lens (a) for a divergent bundle, (b) for a parallel bundle.

contains various colors. We shall discuss each of the monochromatic aberrations first, and afterward take up the chromatic effects.

9.2. Spherical Aberration. Spherical aberration causes a blurring of the image of a point object placed anywhere on the axis. As shown in Fig. 9A(a), the spherical aberration is equal to the interval $M'_1 - M'_6$. Its amount varies with the position of the object point and for any fixed point is approximately proportional to the square of the radius of the zone on the lens surface through which the rays pass. Since many of the lenses in optical instruments are used to focus parallel incident or emergent rays, it is usual for comparison purposes to compute the spherical aberration for parallel incident light. The diagram of Fig. 9A(b) illustrates this special case and shows the position of the paraxial focal point F' as well as the focal points A , B , and C for zones of increasing diameter.

As a measure of the actual magnitudes involved in spherical aberration, we may use the focal lengths for three zones of a lens which were accurately calculated in Table 8II. The results were 9.220 cm for paraxial rays, 9.088 cm for rays traversing a zone of radius $h = 1$ cm, and 8.879 cm for a zone of radius $h = 1.5$ cm. These figures give a spherical aberration of 0.341 cm for the 1.5-cm zone, or about 4 per cent of the paraxial focal length. A graph showing the variation of f with h for this lens is given in Fig. 9B. For small h the curve approximates to a parabola, and since the marginal rays intersect the axis to the left of the paraxial focal point, the spherical aberration is said to be *positive*. A similar curve for an equiconcave lens would bend over to the right, corresponding to *negative* spherical aberration.

A series of positive lenses of the same diameter and paraxial focal length but of different shape is presented in Fig. 9C(a). The alteration of shape represented in this series is known as *bending* the lens. Each lens is labeled by a number q called its *shape factor*, defined by the formula

$$q = \frac{r_2 + r_1}{r_2 - r_1} \quad (9a)$$

As an example, if the two radii of a converging meniscus lens are $r_1 = -15$ cm and $r_2 = -5$ cm, it has a shape factor

$$q = \frac{-5 - 15}{-5 + 15} = -2$$

The usual reason for considering the bending of a lens is to find that shape for which the spherical aberration is a minimum. That such a minimum exists is shown by the graphs of Fig. 9C(b). These curves are drawn for the same lenses as shown in (a), and the values were taken from Table 9I. They were calculated by the ray-tracing methods of Chap. 8, Tables 8I and 8II. It will be noted that lens (5), for which the shape factor q is $+0.5$, has the least spherical aberration. The amount of this aberration for the ray having $h = 1$ cm is shown for the same series of lenses by the curves of Fig. 9D. Over the range of shape factors from about

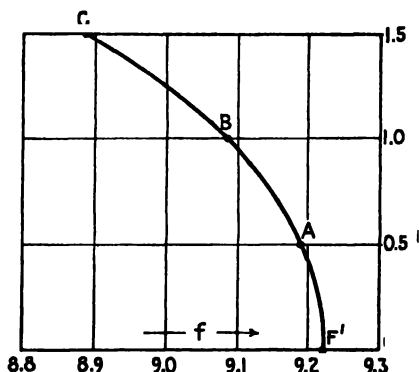


FIG. 9B. Graph showing the variation of focal length with ray height h . The differences of f are a measure of spherical aberration.

$q = +0.4$ to $q = +1.0$ the spherical aberration varies only slightly, since it is close to a minimum. At no point, however, does it go to zero. We therefore see that by choosing the proper radii for the two surfaces of a lens the spherical aberration can be reduced to a minimum but cannot be made to vanish completely.

Reference to the diagrams of Fig. 9.1 will show that with spherical surfaces the marginal rays are deviated through too large an angle. Hence any reduction of this deviation will improve the sharpness of the image.

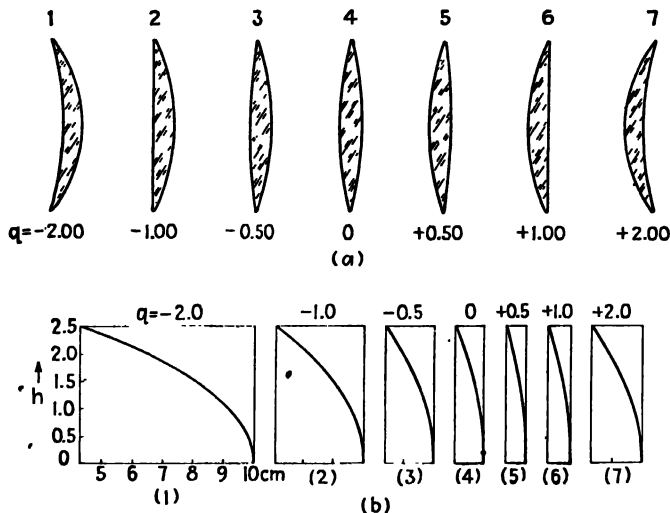


FIG. 9C (a) Lenses of different shapes but with the same power. The difference is one of "bending." (b) Focal length vs. radius for these lenses.

The existence of a condition of minimum deviation in a prism (Sec. 2.9) clearly indicates that when the shape of a lens is changed the deviation of the marginal rays will be least when they enter the first lens surface and leave the second at more or less equal angles. Such an equal division of refraction will yield the smallest spherical aberration. For parallel light incident on a crown-glass lens, this appears from Fig. 9D to occur at a shape factor of about $q = +0.7$, not greatly different from the plano-convex lens, for which $q = +1.0$.

Spherical aberration can be completely eliminated for a single lens by *aspherizing*. This is a tedious hand-polishing process by which various zones of one or both lens surfaces are given different curvatures. For only a few special instruments are such lenses useful enough so that the added expense of hand figuring is justified. Furthermore, since it is

figured for only one object distance, such a lens is not free from spherical aberration for other distances. The most common practice in lens design is to adhere to the simple spherical surfaces and to reduce the spherical aberration by a proper choice of radii.

9.3. Results of Third-order Theory. Although the derivation of an equation for spherical aberration from third-order theory is too lengthy to be given here, some of the resulting equations are of interest. For a thin lens we have the reasonably simple formula

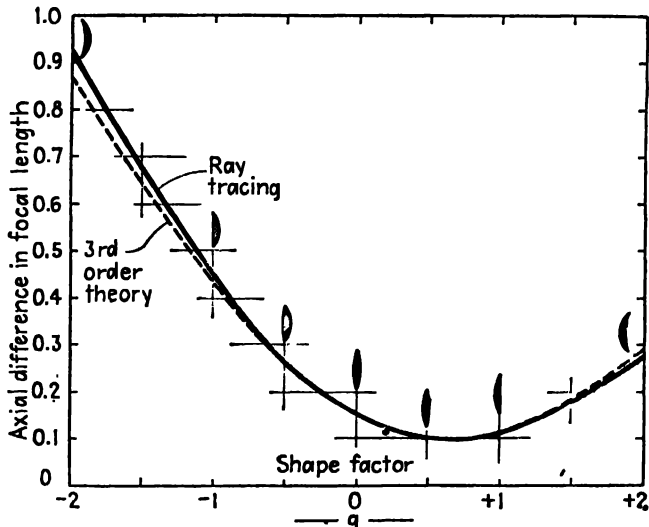


FIG. 9D. Spherical aberration of the ray having $h = 1$ cm, $f = 10$ cm, $d = 2$ cm, $n = 1.517$.

$$\frac{1}{s'_h} - \frac{1}{s'} = \frac{h^2}{8f^3} \cdot \frac{1}{n(n-1)} \left[\frac{n+2}{n-1} q^2 + 4(n+1)pq + (3n+2)(n-1)p^2 + \frac{n^3}{n-1} \right] \quad (9b)$$

where s'_h is the image distance for an oblique ray traversing the lens at a distance h from the axis, s' is the image distance for paraxial rays, and f the paraxial focal length. The constant p is called the *position factor*, and q is the *shape factor* defined above (Eq. 9a). The position factor is defined as

$$p = \frac{s' - s}{s' + s} \quad (9c)$$

Making use of the first-order equation $1/f = (1/s) + (1/s')$, the position factor may also be expressed in terms of f as

$$p = \frac{2f}{s} - 1 = 1 - \frac{2f}{s'} \quad (9d)$$

The difference between the reciprocals of the image distances in Eq. 9b is called the *lateral spherical aberration*,

$$\text{Lat. S.A.} = \frac{1}{s'_h} - \frac{1}{s'}, \quad (9e)$$

while the difference between the two image distances, $s' - s'_h$, is called the *longitudinal spherical aberration*. Solving Eq. 9e for this difference, we obtain

$$\text{Long. S.A.} = s' - s'_h = s's'_h[\text{lat. S.A.}]$$

The image distance for any zone becomes

$$s'_h = \frac{s'}{1 + s'[\text{lat. S.A.}]}$$

A comparison of the third-order theory with the exact results of ray tracing is included in Fig. 9D. When the shape factor is not far from that corresponding to the minimum, the agreement is remarkably good. The numerical results of third-order theory for the seven lenses of Fig. 9C are presented in the last column of Table 9I.

TABLE 9I. SPHERICAL ABERRATION OF LENSES HAVING THE SAME FOCAL LENGTH BUT DIFFERENT SHAPES q
Lens thickness = 1 cm, $f = 10$ cm, $n = 1.5000$, and $h = 1$ cm

Shape of lens	r_1	r_2	q	Ray tracing	Third-order theory
1. Concavo-convex.....	-10.000	- 3.333	-2.00	0.92	0.88
2. Plano-convex.....	∞	- 5.000	-1.00	0.45	0.43
3. Double convex.....	20.000	- 6.666	-0.50	0.26	0.26
4. Equiconvex.....	10.000	-10.000	0	0.15	0.15
5. Double convex.....	6.666	-20.000	+0.50	0.10	0.10
6. Plano-convex.....	5.000	∞	+1.00	0.11	0.11
7. Concavo-convex.....	3.333	10.000	+2.00	0.27	0.29

Equations useful in lens design are obtained by finding the shape factor that will make Eq. 9b a minimum. This may be done by differentiating with respect to the shape factor and equating to zero:

$$\frac{d[\text{lat. S.A.}]}{dq} = \frac{h^2}{8f} \left[\frac{2(n+2)q + 4(n-1)(n+1)p}{n(n-1)^2} \right]$$

Equating to zero and solving for q , one obtains

$$q = -\frac{2(n^2 - 1)p}{n + 2} \quad (9f)$$

as the required relation between shape and position factors to produce minimum spherical aberration. As a rule a lens is designed for some particular pair of object and image distances so that p may be calculated from Eq. 9c. For a lens of a given n the shape factor that will produce a minimum lateral spherical aberration may be obtained at once from Eq. 9f. In order to determine the radii that will correspond to such a calculated shape factor and still yield the proper focal length, one may then use the lens makers' formula

$$\frac{1}{s} + \frac{1}{s'} = (n - 1) \left(\frac{1}{r_1} - \frac{1}{r_2} \right) = \frac{1}{f}$$

Substitution of values of s , s' , and f from Eqs. 9c and 9d gives the following useful set of equations, due to Coddington:

$$\left. \begin{aligned} s &= \frac{2f}{1 + p} & s' &= \frac{2f}{1 - p} \\ r_1 &= \frac{2f(n - 1)}{q + 1} & r_2 &= \frac{2f(n - 1)}{q - 1} \end{aligned} \right\} \quad (9g)$$

The last two relations give the radii in terms of q and f . Division of one of these by the other gives

$$\frac{r_1}{r_2} = \frac{q - 1}{q + 1} \quad (9h)$$

As a problem let us suppose that a single lens is to be made with a focal length of 10 cm and that we wish to find the radii of the surfaces which will give the minimum spherical aberration for parallel incident light. For simplicity we shall assume that the glass has an index $n = 1.50$. In using Eq. 9b the position factor p and the shape factor q must first be determined. Substitution of $s = \infty$ and $s' = 10$ cm in Eq. 9c gives

$$p = \frac{10 - \infty}{10 + \infty} = -1$$

It may be seen that if s is not infinite but is allowed to approach infinity, the ratio $(s' + s):(s' - s)$ will approach the value -1 , and will in the limit be equal to this. Substituting this position factor in Eq. 9f, we obtain

$$q = -\frac{2(2.25 - 1)(-1)}{1.5 + 2} = \frac{2.5}{3.5} = 0.714$$

This value falls at the minimum of the curve of Fig. 9D. The ratio of the two radii is given by Eq. 9h as

$$\frac{r_1}{r_2} = \frac{0.714 - 1}{0.714 + 1} = \frac{-0.286}{1.714} = -0.167$$

The negative sign means that the surfaces curve in opposite directions, and the numerical value indicates a ratio of the radii of about 6:1. Their individual values are found from Eq. 9g to be

$$r_1 = \frac{10}{1.714} = 5.83 \text{ cm} \quad \text{and} \quad r_2 = \frac{10}{0.286} = -35.0 \text{ cm}$$

Such a lens lies between lenses (5) and (6) in Fig. 9C, and has essentially the same amount of spherical aberration as either one. For this reason plano-convex lenses are often employed in optical instruments with the convex side facing the parallel incident rays. Should such a lens be turned around so that the flat side is toward the incident light, its shape factor becomes $q = -1.0$, and the spherical aberration increases about fourfold.

Although spherical aberration cannot be entirely eliminated for a single spherical lens, it is possible to do so for a combination of two or more lenses of opposite sign. The amount of spherical aberration introduced by one lens of such a combination must be equal and opposite to that introduced by the other. If for example the doublet is to have a positive power and no spherical aberration, the positive lens should have the greater power and its shape should be at or near that for minimum spherical aberration, while the negative lens should have a smaller power and its shape should not be near that for the minimum. Neutralization by such an arrangement is possible because spherical aberration varies as the cube of the focal length, and therefore changes sign with the sign of f (see Eq. 9b). In a cemented lens of two elements, the two interfaces should have the same radius. The other two may then be varied and thus used to correct for spherical aberration. With four radii to manipulate, other aberrations like chromatic aberration can be reduced at the same time. This subject will be considered in Sec. 9.10.

9.4. Fifth-order Spherical Aberration. The two curves that were given in Fig. 9D show that, for a lens having a shape factor anywhere near the optimum, the agreement between the exact results of ray tracing and the approximate results of third-order theory is remarkably good. For larger values of h , however, and for shapes further removed from the optimum, appreciable differences occur. This indicates the necessity of including the fifth-order terms in the theory. The third-order equa-

tion 9*b* shows that spherical aberration should be proportional to h^2 , so that the curves in Fig. 9*D* should be parabolas. Nevertheless accurate measurements show that for larger h departures from proportionality to h^2 do occur and that spherical aberration is more closely represented by an equation of the form

$$\text{Long. S.A.} = ah^2 + bh^4 \quad (9i)$$

where a and b are constants. The term ah^2 represents the third-order effect and bh^4 the fifth-order effect. Some numerical results for a single lens, indicating the necessity for the inclusion of the latter term, are shown in Table 9II. The **boldface** values in the fifth row are the true values for longitudinal spherical aberration, obtained by ray-tracing methods, while those in the last row correspond to a parabola that has been fitted at $h = 1.0$ cm to the equation

$$\text{Long. S.A.} = a'h^2$$

with $a' = 0.11530 \text{ cm}^{-1}$.

TABLE 9II. FIFTH-ORDER CORRECTION TO SPHERICAL ABERRATION
 $f = 10 \text{ cm}$, $r_1 = +5 \text{ cm}$, $r_2 = \infty$, $n = 1.5000$, $d = 1 \text{ cm}$

1. h , cm	0.5	1.0	1.5	2.0	2.5	3.0
2. ah^2	0.02839	0.11356	0.25551	0.45424	0.70975	1.02204
3. bh^4	0.00011	0.00174	0.00881	0.02784	0.06797	0.14094
4. $ah^2 + bh^4$	0.02850	0.11530	0.26432	0.48208	0.77772	1.16298
5. Ray tracing.....	0.02897	0.11530	0.26515	0.48208	0.77973	1.16781
6. Parabola.....	0.02882	0.11530	0.25942	0.46120	0.71812	1.03770

The second row gives the third-order corrections ah^2 and the third row the fifth-order corrections bh^4 . The fourth row contains the values calculated from Eq. 9*i* by fitting the curve at the two points $h = 1$ cm and $h = 2$ cm. Assuming the values 0.11530 and 0.48208 at these points, the constants become

$$a = 0.11356 \quad \text{and} \quad b = 0.00174$$

A comparison of the totals in the fourth row with the correct values in the fifth row reveals the excellent agreement of the latter with Eq. 9*i*. Graphs of the values in rows 2 and 3 are given in Fig. 9*E*, and show the negligible contribution of the fifth-order correction at small values of h . If only the third-order aberration were present in a lens it would be possible to combine a positive and a negative lens having equal aberrations to obtain a combination corrected for all zones. Because they actually

would have different amounts of fifth-order aberration, however, such a combination can be corrected for one zone only.

A graph illustrating the spherical aberration of a cemented doublet which is corrected for the marginal zone is shown in Fig. 9F. It will be

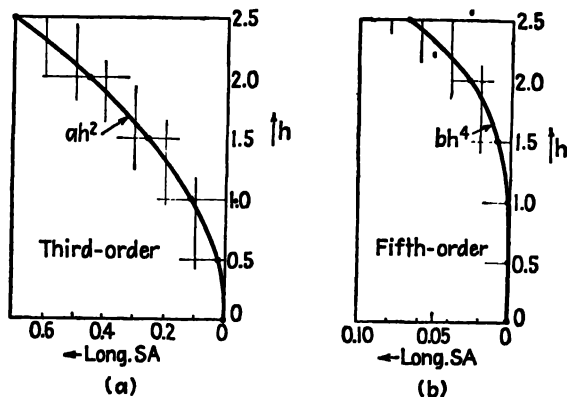


FIG. 9E. Third-order and fifth-order contributions to longitudinal spherical aberration.

seen that the curve comes to zero only at the origin and at the margin. The combination becomes badly overcorrected if the aperture is further increased. The plane of best focus lies a little to the left of the paraxial and marginal focal points, and its position (the vertical broken line) corresponds to that of the circle of least confusion.

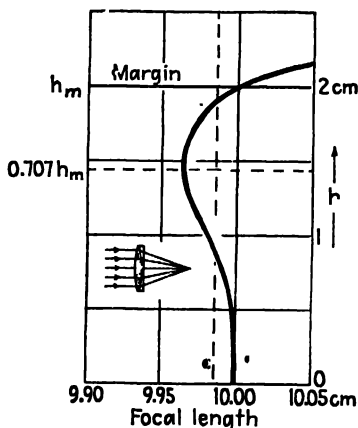


FIG. 9F. Spherical aberration of a corrected doublet as used in telescopes.

Let a and b in Eq. 9i represent the constants for a thin-lens doublet. If the combination is to be corrected at the margin, i.e., for a ray at the height h_m , we must have

$$\text{Long. S.A.} = ah_m^2 + bh_m^4 = 0$$

or

$$a = -bh_m^2$$

Substitution in Eq. 9i yields

$$\text{Long. S.A.} = -bh_m^2h^2 + bh^4$$

where h_m is fixed and h may take any value between 0 and h_m . To find where this expression has a maximum value, we differentiate with respect to h and equate to zero, as follows:

$$\frac{d[\text{long. S.A.}]}{dh} = -2bh_m^2h + 4bh^3 = 0$$

Dividing by $-2bh$, we obtain

$$h = h_m \sqrt{\frac{1}{3}} = 0.707h_m$$

as the radius of the zone at which the aberration reaches a maximum (see Fig. 9F). In lens design spherical aberration is always investigated by tracing a ray through the combination for the zone of radius $0.707h_m$.

9.5. Coma. The second of the monochromatic aberrations of third-order theory is called *coma*. It derives its name from the cometlike appearance of the image of a point object located just off the lens axis.

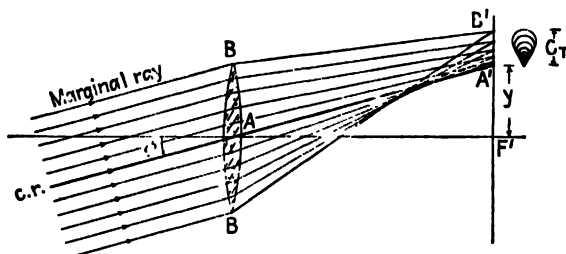


FIG. 9G. Illustrating coma, the second of the five Seidel aberrations of a lens. Only the tangential fan of rays is shown.

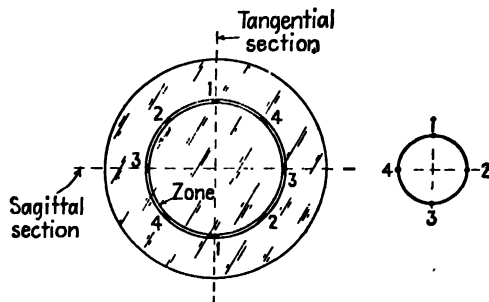


FIG. 9H. Each zone of a lens forms a ring-shaped image called the *comatic circle*.

Although the lens may be corrected for spherical aberration and may bring all rays to a good focus on the axis, the quality of the images of points just off the axis will not be good unless the lens is also corrected for coma. Figure 9G illustrates this lens defect for a single object point infinitely distant and off the axis. Of the fan of rays in the meridian plane that is shown, only those through the center of the lens form an image at A' . Two rays through the margin come together at B' . Thus it appears that the magnification is different for different parts of the lens. If the magnification for the outer rays through a lens is greater than that for the central rays, the coma is said to be *positive*, while if the reverse is true the coma is said to be *negative*.

The shape of the image of an off-axis object point is shown at the upper right in Fig. 9G. Each of the circles represents an image from a different zone of the lens. Details of the formation of the *comatic circle* by the light from one zone of the lens are shown in Fig. 9H. Rays (1), which correspond to the *tangential rays* *B* in Fig. 9G, cross at (1) on the comatic circle, while rays (3), called the *sagittal rays*, cross at the bottom of that circle. In general all points on a comatic circle are formed by the crossing of pairs of rays passing through two diametrically opposite points of the same zone. Third-order theory shows that the radius of a comatic circle is given by

$$C_s = \frac{yh^2}{f^3} (Gp + Wq) \quad (9j)$$

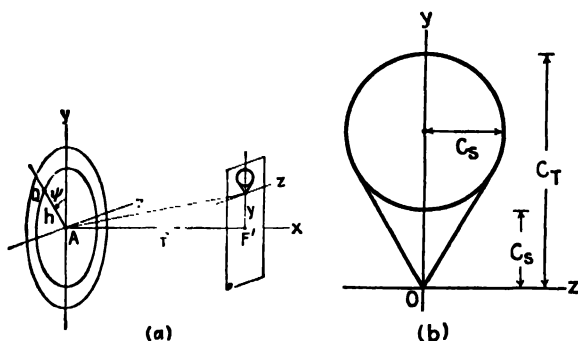


Fig. 9I (a) Geometry of coma, (b) Tangential and sagittal coma.

where y , h , and f are the distances indicated in Fig. 9I(a), and p and q are the Coddington position and shape factors given by Eqs. 9c and 9a. The other two constants are defined as

$$G = \frac{3(2n + 1)}{4n} \quad \text{and} \quad W = \frac{3(n + 1)}{4n(n - 1)}$$

The shape of the comatic figure is given by

$$y = C_s(2 + 2 \cos \psi) \quad z = C_s \sin 2\psi$$

which shows that the tangential coma C_t is three times the sagittal coma C_s . See Fig. 9I (b). Thus

$$C_t = 3C_s$$

The condition required of a lens if it is to be free of coma can be stated in terms of the Seidel sums, namely, that $S_2 = 0$, or more specifically, in terms of the sine condition,

$$ny \sin \theta = n'y' \sin \theta' \quad (8l)$$

Here y and y' are the object and image heights, n and n' the indices of the object and image spaces, and θ and θ' the slope angles of the ray in these two spaces. Now we have seen above that to prevent coma the lateral magnification must be constant for all zones of the lens, or

$$m = \frac{y'}{y} = \text{const.}$$

Solving Eq. 8l for $\sin \theta / \sin \theta'$, we have the requirement that

$$\frac{\sin \theta}{\sin \theta'} = \frac{y'}{y} \cdot \frac{n'}{n} = \text{const.}$$

Any optical system is therefore free of coma if, in the absence of other aberrations, this equation is satisfied for all values of θ . In lens design coma is sometimes tested for by plotting the ratio $\sin \theta : \sin \theta'$ against the height of the incident ray. Because most lenses are used with parallel incident or emergent light, it is customary to replace $\sin \theta$ by h , the height of the ray above the axis, and to write the sine condition in the special form

$$\frac{h}{\sin \theta'} = \text{const.} \quad (9k)$$

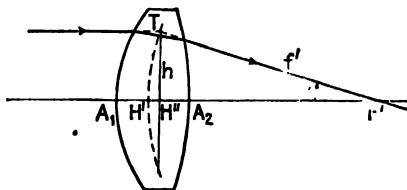


FIG. 9J. For a lens to be free of coma the principal surface of a lens should be spherical.

The ray diagram of Fig. 9J shows that the constant in this equation is the focal distance measured along the image ray, which we shall call f' . To prevent coma, f' must be the same for all values of h . Since freedom from spherical aberration requires that all rays cross the axis at F' , an accompanying freedom from coma requires that the "principal plane" be a spherical surface (represented by the broken line in the figure) of radius f' . It is thus seen that, whereas spherical aberration is concerned with the crossing of the rays at the focal point, coma is concerned with the position of the principal point, *i.e.*, with the shape of the principal surface.

To see how coma is affected by the shape of the lens, a curve of the differences between the paraxial focal lengths $H'F'$ and the distances f' to the principal surface measured along a ray incident at the height $h = 1.0$ cm is plotted in Fig. 9K for the seven lenses of Table 9III. Spherical aberration is present in all these lenses, so that the rays do not come to a focus at the same point on the axis and the sine condition is not a true measure of coma. By subtracting the spherical aberration

from the differences between f' and the paraxial focal lengths $H'F'$ (labeled "sine condition" in Fig. 9K) the bottom curve giving the coma is obtained. The point where the latter line crosses the axis gives the

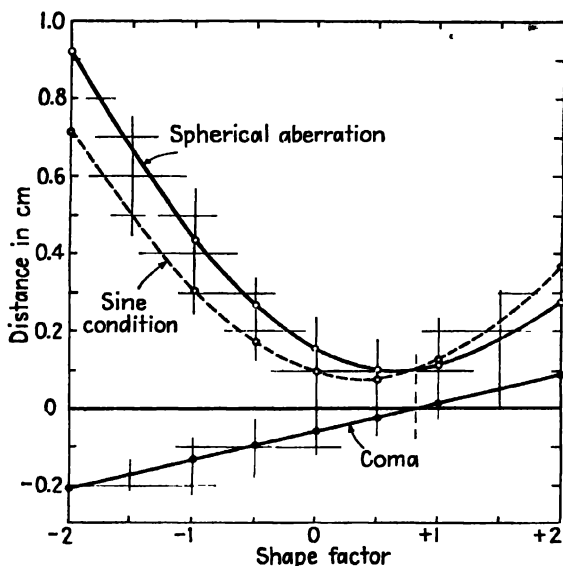


FIG. 9K. Spherical aberration and coma for lenses of different shapes.

lens shape for which there is no coma. It will be noted that this comes very close to the shape for minimum spherical aberration. In other words, a single lens may be made free of coma and at the same time have practically the least possible spherical aberration.

TABLE 9III. SINE CONDITION AND COMA FOR LENSES OF THE SAME FOCAL LENGTH BUT OF DIFFERENT SHAPES

Lens thickness = 2 cm, $f = 10$ cm, $n = 1.51700$, and $h = 1$ cm

Shape of lens	Shape factor	Spherical aberration	$H'F' - f'$	Coma
1. Concavo-convex.....	-2.00	0.92	0.71	-0.21
2. Plano-convex.....	-1.00	0.45	0.30	-0.15
3. Double convex.....	-0.50	0.26	0.16	-0.10
4. Equiconvex.....	0	0.15	0.09	-0.06
5. Double convex.....	+0.50	0.10	0.08	-0.02
6. Plano-convex.....	+1.00	0.11	0.13	+0.02
7. Concavo-convex.....	+2.00	0.27	0.37	+0.10

In order to calculate the value of q that will make Eq. 9j vanish, C_s is set equal to zero. There results

$$q = -\frac{G}{W}p \quad (9l)$$

If the shape and position factors of a single lens obey this relation, the lens is coma-free. A doublet designed to correct for spherical aberration can at the same time be corrected for coma. A graph showing the residual spherical aberration and coma for a telescope objective is given in Fig. 9L.

9.6. Aplanatic Points of a Spherical Surface. An optical system free of both spherical aberration and coma is said to be *aplanatic*. The significance of an aplanatic surface in the simple case of a single surface has already been discussed in Sec. 1.6. An *aplanatic lens* may also be found for any particular pair of conjugate points, although in general it will need to be an aspherical lens. Except for a few special cases, no lens combination with spherical surfaces is completely free of both these aberrations.

One special case which is of considerable importance in microscopy is that of a single spherical refracting surface. To demonstrate the existence of aplanatic points for a single surface, a useful construction, originally discovered by Huygens, will first be described. In Fig. 9M(a) the ray RT represents any ray in the first medium, of index n , incident on the surface at T and making an angle ϕ with the normal NC . Around C as a center and with radii $\rho' = nr/n'$ and $\rho = n'r/n$, the broken circular arcs are drawn as shown. Where RT , when produced, intersects the larger circle a line JC is drawn, and this intersects the smaller circle at K . Then TK gives the direction of the refracted ray in accordance with the law of refraction.* Furthermore any ray whatever directed toward J will be refracted through K .

The aplanatic points of a single surface are located where the two construction circles cross the axis [see Fig. 9M(b)]. All rays initially traveling toward M will pass through M' , and similarly all rays diverging from M' will after refraction appear to originate at M . The application

* For a proof of this proposition, see J. P. C. Southall, "Mirrors, Prisms, and Lenses," 3d ed., p. 512, The Macmillan Company, New York

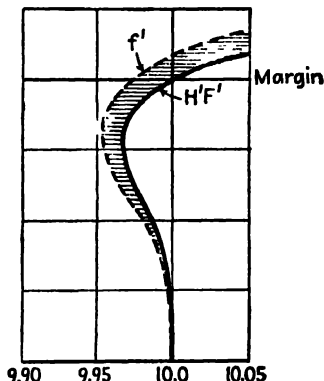


FIG. 9L. Curves plotted for a cemented doublet showing the variable position of the focal point F' (spherical aberration), and the variable focal length f' (coma = $H'F' - f'$).

of this principle to a microscope is illustrated in Fig. 9N. A drop of oil having the same index as the hemispherical lens is placed on the micro-

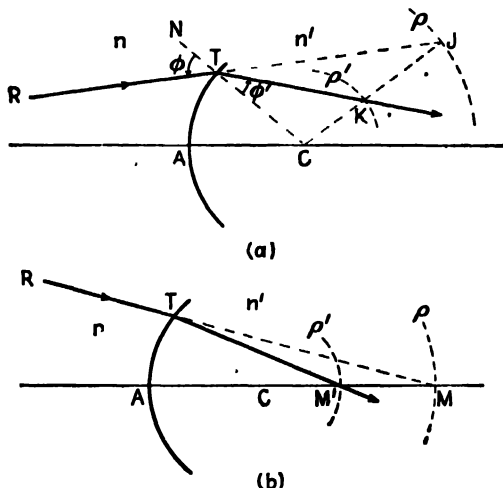


FIG. 9M. (a) Construction for refraction at a spherical surface. $\phi = \sin^{-1}(\rho/n)$, $\phi' = \sin^{-1}(\rho'/n')$. (b) Location of the aplanatic points of a single spherical surface.

scope slide and the lens lowered into contact as shown. All rays from an object at M leave the hemispherical surface after refraction as though they came from M' , and this introduces a lateral magnification of $M'A/MA$. If a second lens is added which has the center of its concave surface at M' (and therefore is normal to all rays), refraction at the upper, convex surface will give added lateral magnification. There is a limit to this process which is set by chromatic aberration (Sec. 9.10).

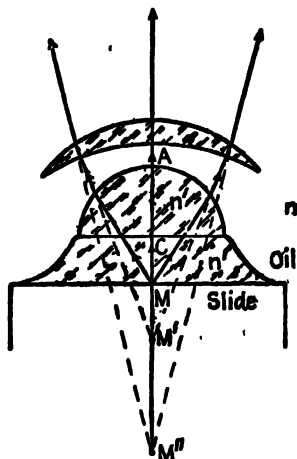


FIG. 9N. Aplanatic surfaces of the first elements oil-immersion microscope objective.

9.7. Astigmatism. If the first two Seidel sums vanish, all rays from points on or very close to the axis of a lens will form point images and there will be no spherical aberration or coma. When the object point lies at some distance away from the axis, however, a point image will be formed only if the third sum S_3 is zero. If the lens fails to satisfy this third condition, it is said to be afflicted with *astigmatism*, and the resulting blurred

images are said to be astigmatic. The formation of a virtual astigmatic image by a plane surface was discussed in Sec. 2.6, and that

of real images from a concave spherical mirror in Sec. 6.9. To help understand the formation of astigmatic images by a lens, a ray diagram has been drawn in perspective in Fig. 90(a). Considering the rays from a point object Q , all those in the fan contained in the vertical or tangential plane cross at T , while the fan of rays in the horizontal or sagittal plane crosses at S . The tangential and sagittal planes intersect the lens in RS and JK respectively. Rays in these planes are chosen because they locate the two focal lines T and S formed by all rays going through the

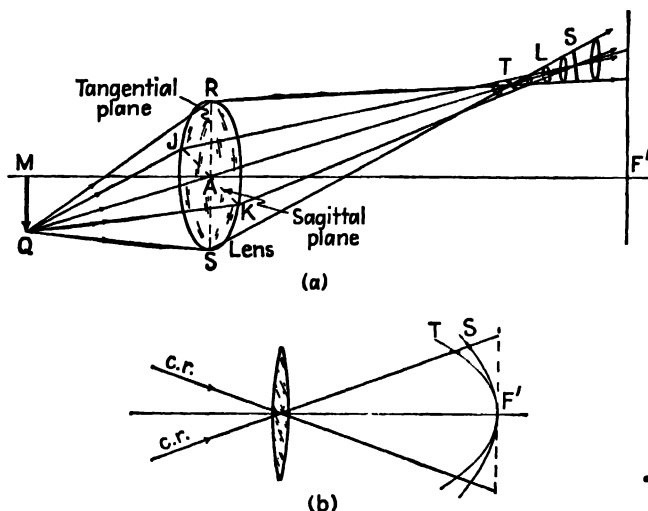


FIG. 90. (a) Perspective diagram showing the two focal lines which constitute the image of an off-axis object point Q . (b) Loci of the tangential and sagittal images. The two surfaces are paraboloids of revolution.

lens. These are perpendicular to their respective tangential and sagittal planes. At L the image is approximately disk-shaped, and constitutes the circle of least confusion for this case.

If the positions of the T and S images are determined for a wide field of distant object points, their loci will form paraboloidal surfaces whose sections are shown in Fig. 90(b). The amount of astigmatism, or *astigmatic difference*, for any pencil of rays is given by the distance between these two surfaces measured along the chief ray. On the axis, where the two surfaces come together, the astigmatic difference is zero; away from the axis it increases approximately as the square of the image height. Astigmatism is said to be positive when the T surface lies to the left of S , as shown in the diagram. It should be noted that for a concave mirror, Fig. 6*N*, the sagittal surface is a plane coinciding with the paraxial focal plane.

If, as in Fig. 9P, the object is a spoked wheel in a plane perpendicular to the axis with its center at M , the rim would be found to be in focus on the T surface while the spokes would be in focus on the S surface. It is for this reason that the terms "tangential" and "sagittal" are applied to the planes and images. On the surface T all images will be lines parallel to the rim as shown at the left in Fig. 9P, while on the surface S all images will be lines parallel to the spokes as shown at the right.

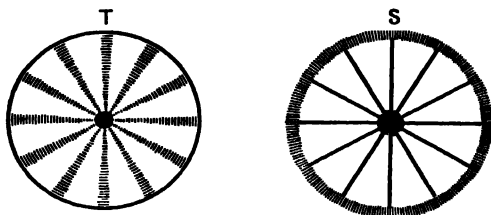


FIG. 9P. Astigmatic images of a spoked wheel.

Equations giving the astigmatic image distances for a single refracting surface are*

$$\frac{n \cos^2 \phi}{s} + \frac{n' \cos^2 \phi'}{s'_t} = \frac{n' \cos \phi'}{r} - \frac{n \cos \phi}{r} \quad (9m)$$

$$\frac{n}{s} + \frac{n'}{s'_s} = \frac{n' \cos \phi'}{r} - \frac{n \cos \phi}{r}$$

where ϕ and ϕ' are the angles of incidence and refraction of the chief ray, r the radius of curvature, s the object distance, and s'_t and s'_s the T and S image distances, the latter being measured along the chief ray. For a spherical mirror these equations reduce to

$$\frac{1}{s} + \frac{1}{s'_t} = \frac{1}{f \cos \phi} \quad \text{and} \quad \frac{1}{s} + \frac{1}{s'_s} = \frac{\cos \phi}{f}$$

Coddington has shown that for a thin lens in air with an aperture stop at the lens, the positions of the tangential and sagittal images are given by

$$\frac{1}{s} + \frac{1}{s'_t} = \frac{1}{\cos \phi} \left(\frac{n \cos \phi'}{\cos \phi} - 1 \right) \left(\frac{1}{r_1} - \frac{1}{r_2} \right) \quad (9n)$$

$$\frac{1}{s} + \frac{1}{s'_s} = \cos \phi \left(\frac{n \cos \phi'}{\cos \phi} - 1 \right) \left(\frac{1}{r_1} - \frac{1}{r_2} \right)$$

The angle ϕ is the angle of obliquity of the incident chief rays, and ϕ' the angle of this ray within the lens. Therefore $n = \sin \phi / \sin \phi'$. The

* For a derivation of these formulas see G. S. Monk, "Light, Principles and Experiments," 1st ed., p. 424, McGraw-Hill Book Company, Inc., New York,

application of these formulas to thin lenses shows that the astigmatism is approximately proportional to the focal length and is very little improved by changing the shape.

Although a contact doublet composed of one positive and one negative lens shows considerable astigmatism, the introduction of another element consisting of a stop or a lens can be made to greatly reduce it. By the proper spacing of the lens elements of any optical system, or by the proper location of a stop if one is used, the curvature of the astigmatic image surfaces can be changed considerably. Four important stages in the flattening of the astigmatic surfaces due to these alterations are shown in Fig. 9Q. Diagram (a) represents the normal shape of the T and S surfaces for a contact doublet or a single lens. In diagram (b) the separation of lens elements is such that the two surfaces fall together at P . Further alteration of the lens shapes and their spacing may be made

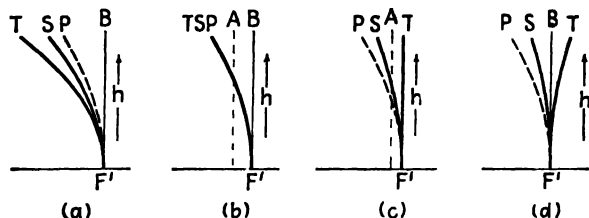


FIG. 9Q. Diagrams showing the astigmatic surfaces T and S in relation to the fixed Petzval surfaces P , as the spacing between lenses (or between lens and stop) is changed.

and the T and S curves straightened, as in diagram (c), or moved still farther apart until they are bisected by the normal plane through the focal point F' , as in diagram (d). Of these four arrangements, only the second is free of astigmatism. The single paraboloidal surface P , over which point images are formed, is called the *Petzval surface*.

9.8. Curvature of Field. If for an optical system the first three Seidel sums are zero, the system will form point images of point objects on as well as off the axis. Under these circumstances the images fall on the curved Petzval surface where the tangential and sagittal surfaces come together, as in Fig. 9Q(b). Even though astigmatism is corrected for such a system, the focal surface is curved. If a flat screen is placed in position B , the center of the field will be in sharp focus but the edges will be quite blurred. With a screen at A , the center of the field and the field margins will be blurred, while sharp focus will be obtained about halfway out.

Mathematically a Petzval surface exists for every optical system, and if the powers and refractive indices of the lenses remain fixed the shape of the Petzval surface cannot be changed by altering the shape factors

of the lenses or their spacing. Such alterations, however, will change the shapes of the T and S surfaces, but always in such a way that the ratio of the distances PT and PS is 3:1. It will be noted that this ratio is maintained throughout Fig. 9Q. If a system is designed to make the T surface flat, as in Fig. 9Q(c), the 3:1 ratio of distances requires the S surface to be curved, but not strongly so. If a screen is placed at a compromise position A , the images over the entire field will be in reasonably good focus. This condition of correction is commonly used for certain types of photographic lenses. If more negative astigmatism is introduced the condition shown in Fig. 9Q(d) is reached, in which the T surface is convex and the S surface is concave by an equal amount. In this case a screen placed at the paraxial focus will show considerable blurring at the field edges.

Curvature of field may be corrected for a single lens by means of a stop. Acting as a second element of the system, a stop limits the rays

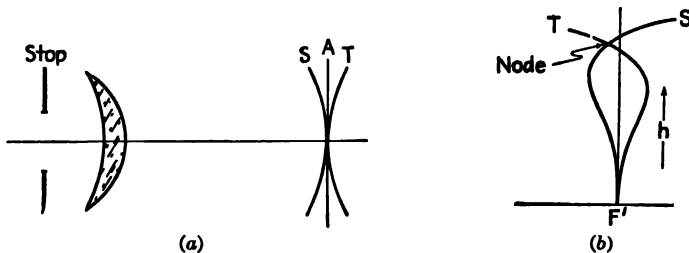


FIG. 9R. (a) A properly located stop may be used to reduce field curvature. (b) Astigmatic surfaces for an "anastigmat" camera lens.

from each object point in such a way that the paths of the chief rays from different points go through different parts of the lens [Fig. 9R(a)]. Certain manufacturers of inexpensive box cameras employ a single meniscus lens and a stop and with them obtain reasonably good imagery. The stop is located in front of the lens, with the light incident on the concave surface. Although the compromise field is flat and sharp focus is obtained at the center, astigmatism gives rise to blurred images at the margins.

In complex lens systems it is possible, because of differences in third- and fifth-order corrections, to control the astigmatism and cause the tangential and sagittal surfaces to come together at an outer zone as well as at the center of the field. Typical curves for the camera objective called an "anastigmat" are shown in Fig. 9R(b). Experience has shown that the best state of correction is obtained by making the crossover point, called the node, occur at a relatively short distance in front of the focal plane.

9.9. Distortion. Even though an optical system were designed so that the first four Seidel sums were zero, it could still be affected by the fifth aberration known as *distortion*. To be free of distortion a system must have uniform lateral magnification over its entire field. A pinhole camera is ideal in this respect for it shows no distortion; all straight lines connecting each pair of conjugate points in the object and image planes pass through the opening. Constant magnification for a pinhole camera as well as for a lens implies, as may be seen from Fig. 9S, that

$$\frac{\tan \phi'}{\tan \phi} = \text{const.}$$

The common forms of image distortion produced by lenses are illustrated in Fig. 9T. Diagram (a) represents the undistorted image of an object consisting of a rectangular wire mesh. The second diagram shows *barrel* distortion, which arises when the magnification decreases towards the edge of the field. The third diagram represents *pincushion* distortion, corresponding to a greater magnification at the borders.

A single thin lens is practically free of distortion for all object distances. It cannot, however, be free of all the other aberrations at the same time. If a stop is placed in front of or behind a thin lens, distortion is invari-

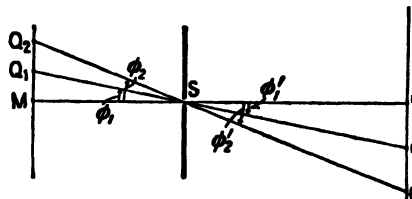


FIG. 9S. A pinhole camera shows no distortion.

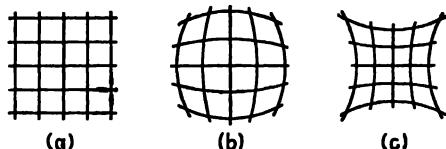


FIG. 9T. Images of a rectangular wire screen shown (a) with no distortion, (b) with barrel distortion, and (c) with pincushion distortion.

ably introduced; if it is placed at the lens, there is no distortion. Frequently in the design of good camera lenses astigmatism, as well as distortion, is corrected for by a nearly symmetrical arrangement of two lens elements with a stop between them.

To illustrate the principles involved, consider the lens shown in Fig. 9U(a), which has a front stop. Rays from object points like *M*, at or near the axis, go through the central part of the lens, while rays from off-axis object points like *Q*₂ are refracted only by the upper half. In effect the stop decreases the ratio of image to object distances, thereby reducing the lateral magnification below that obtaining for object points

near the axis. This system therefore suffers from barrel distortion. When the lens and stop are turned around, as in Fig. 9U(b), the ratio of image to object distances is seen to increase as the object point lies farther off the axis. The result is increased magnification and pin-cushion distortion.

By combining two identical lenses with a stop midway between them as in Fig. 9U(c), a system is obtained which because of its symmetry is free from distortion for unit magnification. With other magnifications

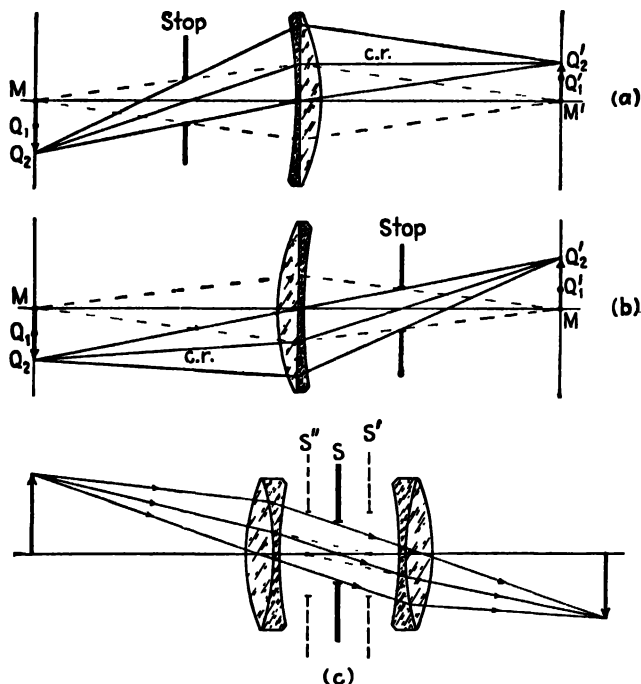


Fig. 9U. (a) The stop in front of the lens gives rise to barrel distortion. (b) Located behind, it gives pincushion distortion. (c) A symmetrical doublet is relatively free of distortion.

however the lenses must be corrected for spherical aberration with respect to the entrance and exit pupils. These two pupils S' and S'' coincide with the principal planes of the combination. Such a corrected lens system is called an *orthoscopic doublet*, or rapid rectilinear lens. Because this combination cannot be corrected for spherical aberration for the object and image planes and for the entrance and exit pupils at the same time, the lens suffers from this aberration as well as from astigmatism. Photographic lenses of this type are discussed in Sec. 10.3.

Summarizing very briefly the various methods of correcting for aberrations

tions, spherical aberration and coma can be corrected by using a contact doublet of the proper shape; astigmatism and curvature of field require for their correction the use of several separated components; and distortion may be minimized by the proper placement of a stop.

9.10. Chromatic Aberration. In the discussion of the third-order theory given in the preceding sections, no account has been taken of the change of refractive index with color. The assumption that n is constant amounts to investigating the behavior of the lens for monochromatic light only. Because the refractive index of all transparent media varies with color, a single lens forms not only one image of an object

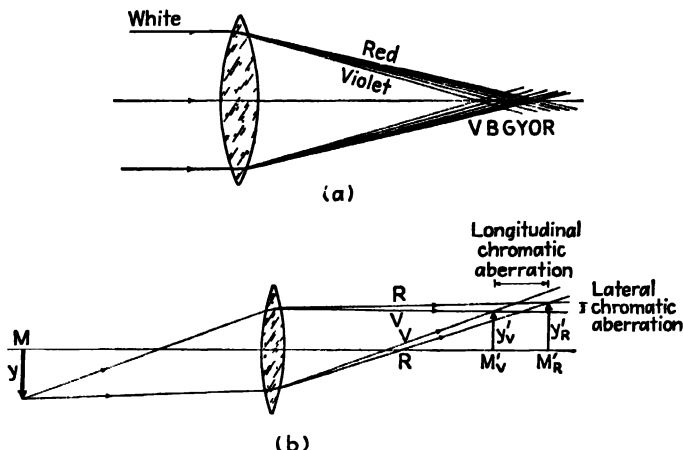


FIG. 9V. (a) Chromatic aberration of a single lens for parallel light. (b) The two types of chromatic aberration of an image.

but a series of images, one for each color of light present in the beam. Such a series of colored images of an infinitely distant object point on the axis of the lens is represented diagrammatically in Fig. 9V(a). The prismatic action of the lens, which increases toward its edge, is such as to cause dispersion and to bring the violet light to a focus nearest to the lens.

As a consequence of the variation of focal length of a lens with color, the magnification must vary as well. This may be seen by the diagram of Fig. 9V(b), which shows only the red and violet images of an off-axis object point. The horizontal distance between the images is called *axial* or *longitudinal chromatic aberration*, while the vertical difference in height is called *lateral chromatic aberration*. Because these aberrations are often comparable in magnitude with the Seidel aberrations, correction for both lateral and longitudinal color is of considerable importance. As

an indication of relative magnitudes, the longitudinal chromatic aberration of an equiconvex lens of spectacle crown glass having a focal length of 10 cm and a diameter of 3 cm is exactly the same (2.5 mm) as the spherical aberration of marginal rays in the same lens.

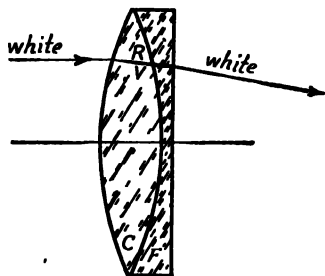


FIG. 9W. An achromatic lens composed of crown- and flint-glass elements.

The dispersion is neutralized, thereby bringing all colors to approximately the same focus. The principle is therefore the same as that of the achromatic prism described in Sec. 2.12. The possibility of achromatizing such a combination rests upon the fact that the dispersions produced by differ-

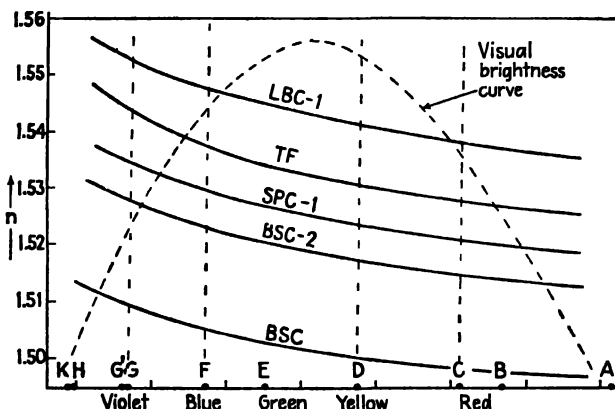


FIG. 9X. Graphs of the refractive indices of several kinds of optical glass. These are called *dispersion curves*.

ent kinds of glass are not proportional to the deviations they produce (Sec. 1.11). In other words, the dispersive powers $1/\nu$ differ for different materials.

Typical dispersion curves showing the variation of n with color are plotted for a number of common optical glasses in Fig. 9X, and the actual

values of the index n for the different Fraunhofer lines (Sec. 1.11) are presented in Table 9IV. The peak of the visual brightness curve* in Fig. 9X occurs not far from the yellow D line. It is for this reason that the index n_D has been chosen by optical designers as the basic index for ray tracing and for the specification of focal lengths. Two other indices, one on either side of n_D , are then chosen for purposes of achromatization. As indicated in the table, the ones most often used are n_C for the red end of the spectrum and n_F or n_G for the blue end.

TABLE 9IV. REFRACTIVE INDICES OF TYPICAL OPTICAL MEDIA FOR FOUR COLORS

Medium	Symbol	I.C.T. type	ν	n_C	n_D	n_F	n_G
Borosilicate crown.....	BSC	500/665	66.5	1.49776	1.50000	1.50529	1.50937
Borosilicate crown.....	BSC-2	517/645	64.5	1.51462	1.51700	1.52264	1.52708
Spectacle crown.....	SPC-1	523/586	58.8	1.52042	1.52300	1.52933	1.53435
Light barium crown.....	LBC-1	541/599	59.7	1.53828	1.54100	1.54735	1.55249
Telescope flint.....	TF	530/516	51.6	1.52762	1.53050	1.53790	1.54379
Dense barium flint.....	DBF	670/475	47.5	1.66650	1.67050	1.68059	1.68882
Light flint.....	LF	576/412	41.2	1.57208	1.57600	1.58606	1.59441
Dense flint.....	DF-2	617/366	36.6	1.61216	1.61700	1.62901	1.63923
Dense flint.....	DF-4	649/338	33.9	1.64357	1.64900	1.66270	1.67456
Extra dense flint.....	EDF-3	720/291	29.1	1.71303	1.72000	1.73780	1.75324
Fused quartz.....	SiO ₂		67.9		1.4585	.	.
Crystal quartz (O ray)....	SiO ₂		70.0		1.5443		
Fluorite.....	CaF ₂		95.4		1.4338	.	

For two thin lenses in contact, the resultant focal length f_D or power P_D of the combination for the D line is given by Eqs. 3j and 3k:

$$\frac{1}{f_D} = \frac{1}{f'_D} + \frac{1}{f''_D} \quad \text{or} \quad P_D = P'_D + P''_D \quad (9o)$$

where the index D indicates that the quantity depends on n_D , f'_D and P'_D refer to the focal length and power of the crown-glass component, and f''_D and P''_D to the focal length and power of the flint-glass component. In terms of indices of refraction and radii of curvature, the power form of the equation becomes

$$P_D = (n'_D - 1) \left(\frac{1}{r'_1} - \frac{1}{r'_2} \right) + (n''_D - 1) \left(\frac{1}{r''_1} - \frac{1}{r''_2} \right) \quad (9p)$$

* Brightness is a sensory magnitude in light just as loudness is a sensory magnitude in sound. Over a considerable range both vary as the logarithm of the energy. The curve shown represents the logarithms of the "standard luminosity curve."

For convenience let

$$K' = \left(\frac{1}{r_1'} - \frac{1}{r_2'} \right) \quad \text{and} \quad K'' = \left(\frac{1}{r_1''} - \frac{1}{r_2''} \right) \quad (9q)$$

Then Eq. (9p) can be more simply written as

$$P_D = (n_D' - 1)K' + (n_D'' - 1)K'' \quad (9r)$$

Similarly, for any other colors or wavelengths like the F and C spectrum lines, we may write

$$\left. \begin{aligned} P_F &= (n_F' - 1)K' + (n_F'' - 1)K'' \\ P_C &= (n_C' - 1)K' + (n_C'' - 1)K'' \end{aligned} \right\} \quad (9r')$$

To make the combination achromatic we make the resultant focal length the same for F and C light. This means, making $P_F = P_C$,

$$(n_F' - 1)K' + (n_F'' - 1)K'' = (n_C' - 1)K' + (n_C'' - 1)K''$$

Multiplying out and canceling, this becomes

$$\frac{K'}{K''} = - \frac{n_F'' - n_C''}{n_F' - n_C'} \quad (9s)$$

Since both the numerator and denominator on the right have positive values, the minus sign shows that one K must be negative and the other positive. This means that one lens must be negative.

Now for the D line of the spectrum the separate powers of the two thin lenses are given by

$$P_D' = (n_D' - 1)K' \quad \text{and} \quad P_D'' = (n_D'' - 1)K'' \quad (9t)$$

Dividing one by the other, this gives

$$\frac{K'}{K''} = \frac{(n_D'' - 1)P_D'}{(n_D' - 1)P_D''} \quad (9t')$$

Equating Eqs. 9s and 9t' and solving for P_D''/P_D' gives

$$\frac{P_D''}{P_D'} = - \frac{(n_D'' - 1)}{(n_F'' - n_C'')} \div \frac{(n_D' - 1)}{(n_F' - n_C')} = - \frac{\nu''}{\nu'} \quad (9u)$$

where ν' and ν'' give the dispersion constants of the two glasses.

These constants, usually supplied by manufacturers when optical glass is purchased, are, according to their definition in Sec. 1.11,

$$\nu' = \frac{n_D' - 1}{n_F' - n_C'} \quad \text{and} \quad \nu'' = \frac{n_D'' - 1}{n_F'' - n_C''} \quad (9v)$$

Values of ν for several common types of glass are given in Table 9IV. Since the dispersive powers are all positive, the negative sign in Eq. 9u indicates that the powers of the two lenses must be of opposite sign. In other words, if one lens is converging the other must be diverging. From the extreme members of Eq. 9u, we obtain

$$\frac{P'_D}{\nu'} + \frac{P''_D}{\nu''} = 0 \quad \text{or} \quad \nu' f' + \nu'' f'' = 0 \quad (9w)$$

Substituting the value of P'_D or that of P''_D from Eq. 9u in Eq. 9w, we obtain

$$P'_D = P_D \left(\frac{\nu'}{\nu' - \nu''} \right) \quad \text{and} \quad P''_D = -P_D \left(\frac{\nu''}{\nu' - \nu''} \right) \quad (9x)$$

The use of the above formulas to calculate the radii for a desired achromatic lens involves the following steps:

1. A focal length f_D and a power P_D are specified.
2. The types of crown and flint glass to be used are selected.
3. If they are not already known, the dispersion constants ν' and ν'' are calculated from Eqs. 9v.
4. P'_D and P''_D are calculated from Eq. 9x.
5. The values of K' and K'' are determined by Eq. 9t.
6. The radii are then found from Eq. 9q.

Calculation 6 is usually made with other aberrations in mind. • •

Example: An achromatic lens having a focal length of 10 cm is to be made as a cemented doublet using crown and flint glasses having the following indices:

Glass	n_C	n_D	n_F	n_G
1. Crown.....	1.50868	1.51100	1.51673	1.52121
2. Flint.....	1.61611	1.62100	1.63327	1.64369

Find the radii of curvature for both lenses if the crown-glass lens is to be equiconvex and the combination is to be corrected for the C and F lines.

Solution: The focal length of 10 cm is equivalent to a power of +10 D. The dispersion constants ν' and ν'' are, from Eq. 9v,

$$\nu' = \frac{1.51100 - 1.00000}{1.51673 - 1.50868} = 63.4783$$

$$\nu'' = \frac{1.62100 - 1.00000}{1.63327 - 1.61611} = 36.1888$$

Applying Eq. 9*x*, we find that the powers of the two lenses must be

$$P'_D = 10 \frac{63.4783}{63.4783 - 36.1888} = +23.2611 \text{ D}$$

$$P''_D = -10 \frac{36.1888}{63.4783 - 36.1888} = -13.2611 \text{ D}$$

The fact that the sum of these two powers $P_D = +10.0000 \text{ D}$ serves as a check on the calculations to this point. Knowing the power required in each lens, we are now free to choose any pair of radii that will give such a power. If two or more surfaces can be made to have the same radius, the necessary number of grinding and polishing tools will be reduced. For this reason the positive element is often made equiconvex, as it is here. Letting $r'_1 = -r'_2$, we apply Eq. 9*q* and then Eq. 9*t* to obtain

$$K' = \frac{1}{r'_1} - \frac{1}{r'_2} = \frac{2}{r'_1} = \frac{P'_D}{n'_D - 1} = \frac{23.2611}{0.51100} = 45.5207$$

from which

$$r'_1 = 0.0439361 \text{ m} = 4.39361 \text{ cm}$$

Since the lens is to be cemented, one surface of the negative lens must fit a surface of the positive lens. This leaves the radius of the last surface to be adjusted to give the proper power of -13.261 D . Therefore we let $r''_1 = -r'_1$, and apply Eqs. 9*t* and 9*q* as before, to find

$$K'' = \frac{1}{r''_1} - \frac{1}{r''_2} = -\frac{1}{0.0439361} - \frac{1}{r''_2} = \frac{P''_D}{n''_D - 1} = \frac{-13.2611}{0.62100} = -21.3544$$

This gives

$$\frac{1}{r''_2} = 21.3544 - \frac{1}{0.0439361} = 21.3544 - 22.7603$$

and

$$r''_2 = -1.4059 \text{ m} = -140.59 \text{ cm}$$

The required radii are therefore

$$\begin{array}{ll} r'_1 = 4.39 \text{ cm} & r''_1 = -4.39 \text{ cm} \\ r'_2 = -4.39 \text{ cm} & r''_2 = -140 \text{ cm} \end{array}$$

It will be noted that, with the crown-glass element of this achromat placed toward incident parallel light, the two exposed surfaces are close to what they should be for minimum spherical aberration and coma. This emphasizes the importance of choosing glasses having the proper dispersive powers.

To see how well this lens has been achromatized, we now calculate its focal length for the four colors corresponding to the C, D, F, and G' lines. By Eq. 9r',

$$\begin{aligned} P_c &= (n'_c - 1)K' + (n''_c - 1)K'' \\ &= 0.50868 \times 45.5207 + 0.61611 (-21.3544) \\ &= 23.1555 - 13.1567 \end{aligned}$$

giving

$$f_c = 10.0012 \text{ cm}$$

Similarly for the colors corresponding to the F and G' lines we obtain

$$\begin{array}{ll} P_F = +9.9988 \text{ D} & \text{or} \quad f_F = 10.0012 \text{ cm} \\ P_{G'} = +9.9804 \text{ D} & \text{or} \quad f_{G'} = 10.0196 \text{ cm} \end{array}$$

The differences between f_c , f_D , and f_F are negligibly small, but $f_{G'}$ is about $\frac{1}{8}$ mm larger than the others. This difference for light outside the region of the C and F lines results in a small circular zone of color about each image point which is called the *secondary spectrum*.

Although the lens in our example would appear to have been corrected for longitudinal chromatic aberration, it has actually been corrected for lateral chromatic aberration. Equal focal lengths for different colors will produce equal magnification, but the different colored images along the axis will coincide only if the principal points also coincide. Practically speaking, the principal points of a thin lens are so close together that both types of chromatic aberration can be assumed to have been corrected by the above arrangement. In a thick lens, however, longitudinal chromatic aberration is absent if the colors corrected for come together at the same axial image point as shown in Fig. 9Y. Because the principal points for blue and red H'_b and H'_r do not coincide, the focal lengths are not equal and the magnification is different for different colors. Consequently the images formed in different colors will have different sizes. This is the lateral chromatic aberration or lateral color mentioned at the beginning of this section.

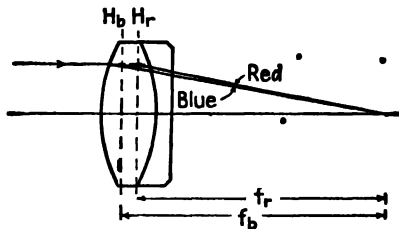


FIG. 9Y. Illustrating how a doublet corrected for longitudinal chromatic aberration is not free of lateral chromatic aberration (greatly exaggerated).

9.11. Separated Doublet. Another method of obtaining an achromatic system is to employ two thin lenses made of the same glass and separated by a distance equal to half the sum of their focal lengths. To see why

this is true we begin with the thick-lens formula, Eq. 5d, as applied to two thin lenses separated by a distance c :

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{c}{f_1 f_2} \quad \text{or} \quad P = P_1 + P_2 - cP_1 P_2 \quad (9y)$$

which, by analogy with Eq. 9r, may be written

$$P = (n_1 - 1)K' + (n_2 - 1)K'' - c(n_1 - 1)(n_2 - 1)K'K'' \quad (9y')$$

The subscripts 1 and 2 are used here in place of the primes to designate the two lenses, and the K 's are given by Eq. 9q. Since the two lenses are of the same kind of glass, we set $n_1 = n_2$, so that

$$P = (n - 1)(K' + K'') - c(n - 1)^2 K'K''$$

If this power is to be independent of the variation of n with color, dP/dn must vanish. This gives

$$\frac{dP}{dn} = K' + K'' - 2c(n - 1)K'K'' = 0$$

Multiplying by $n - 1$ and substituting for each $(n - 1)K$ the corresponding P , we find

$$P_1 + P_2 - 2cP_1 P_2 = 0$$

or

$$c = \frac{P_1 + P_2}{2P_1 P_2} \quad \text{and} \quad c = \frac{f_1 + f_2}{2} \quad (9z)$$

This proves the proposition stated above that two lenses made of the same glass separated by half the sum of their focal lengths have the same focal length for all colors near those for which f_1 and f_2 are calculated. For visual instruments this color is chosen to be at the peak of the visual brightness curve (Fig. 9X). Such spaced doublets are used as oculars in many optical instruments because the lateral chromatic aberration is highly corrected through constancy of the focal length. The longitudinal color, however, is relatively large, due to wide differences in the principal points for different colors.

We have seen in this chapter that a lens may be affected by as many as seven primary aberrations—five monochromatic aberrations of the third and higher orders, and two chromatic aberrations. One might therefore wonder how it is possible to make a good lens at all when rarely can a single aberration be eliminated completely, much less all of them simultaneously. Good usable lenses are nevertheless made by the proper balancing of the various aberrations. The design is guided by the pur-

pose for which the lens is to be used. In a telescope objective, for example, correction for chromatic aberration, spherical aberration, and coma are of primary importance. On the other hand astigmatism, curvature of field, and distortion are not as serious because the field over which the objective is to be used is relatively small. For a good camera lens of wide aperture and field, the situation is almost exactly reversed.

Other treatments of the subject of aberration will be found in the following texts:

"The Principles of Optics," by A. C. Hardy and F. H. Perrin.

"Light, Principles and Experiments," by G. S. Monk.

"Fundamentals of Optical Engineering," by D. H. Jacobs.

"Applied Optics and Optical Design," by A. E. Conrady.

"Technical Optics," by L. C. Martin.

"A Treatise on Reflexion and Refraction," by H. Coddington.

"A System of Applied Optics," by H. D. Taylor.

Problems

1. A thin lens is to be made of glass of index $n = 1.60$, and it is to have a minimum lateral spherical aberration when the object is 60 cm in front of the lens, the real image then being 20 cm in back of the lens. Determine (a) the position factor, (b) the shape factor, (c) the focal length of the lens, and (d) the radii of curvature of the two surfaces.
2. A lens is to be made of borosilicate crown glass of index 1.50 and is to have a focal length of 5 cm. An object is to be located 30 cm in front of the lens. Determine (a) the image distance, and (b) the position factor. If the lens is to have a minimum lateral spherical aberration for these object and image distances, find (c) the shape factor, and (d) the radii of curvature.
3. A thin lens is to be made of glass of index 1.70 and is to have a minimum lateral spherical aberration for distant objects. If the focal length is to be 5 cm, find (a) the position factor, (b) the shape factor, and (c) the radii of curvature of the two surfaces.
4. Calculate the shape factor and the radii of curvature for the lens of Prob. 1 if it is to have no coma.
5. Calculate the shape factor and the radii of curvature for the lens of Prob. 2 if it is to have no coma.
6. The end of a glass rod of index 1.60 is ground and polished with a convex spherical surface of radius 6 cm. Find the aplanatic points of this surface in air.
7. A meniscus lens 1 cm thick, and of index 1.50, is to be aplanatic for two points located on the concave side of the lens. If the nearer point is to be 5 cm from the concave surface, find (a) the radii of the two lens surfaces, and (b) the distance to the farther point. (NOTE: Both points are in air.)
8. A meniscus lens 1 cm thick, and of index 1.60, is to be aplanatic for two points 4 cm apart. Calculate (a) the two radii of curvature and (b) the distances from the concave surface to the two points.
9. An achromatic lens with a focal length of 25 cm is to be made of crown and flint glasses of the types BSC-2 and DF-2 (Table 9IV). If the crown-glass lens is to be

equiconvex and the combination is to be cemented, what must be the radii of curvature to correct for the C and F lines?

10. An achromatic lens with a focal length of 10 cm is to be made of crown and flint glasses of the types SPC-1 and DF-4 (see Table 9IV). If the flint lens is to have its outer face plane and the combination is to be cemented, find the radii of the other three surfaces required to accomplish correction for the C and F lines.

11. Calculate the focal length of the lens in Prob. 9 for the C, D, F, and G' lines.

12. Calculate the focal length of the lens in Prob. 10 for the C, D, F, and G' lines.

CHAPTER 10

OPTICAL INSTRUMENTS

The design of efficient optical instruments is the ultimate purpose of geometrical optics. The principles governing the formation of images by a single lens, and occasionally by simple combinations of lenses, have been set forth in the previous chapters. These principles find a wide variety of applications in the many practical combinations of lenses, frequently including also mirrors or prisms, which fall in the category of optical instruments. This subject is one of such large scope, and has developed so many ramifications, that in a book devoted to the funda-

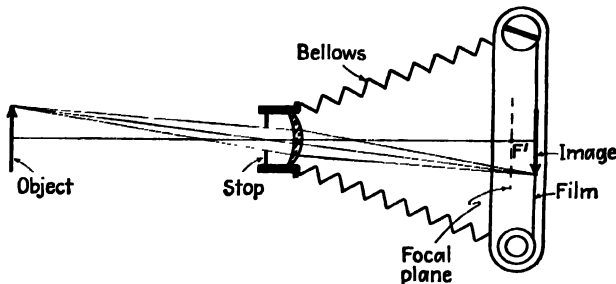


FIG. 10A. Principle of a camera.

mentals of optics it is only possible to describe the principles involved in a few standard types of instrument. In this chapter a description will be given of the more important features of camera lenses, magnifiers, microscopes, telescopes, and oculars. These will serve to illustrate some applications of the basic ideas already discussed and will, it is hoped, be of interest to the student who has used, or expects to use, some of these instruments.

10.1. Photographic Objectives. The fundamental principle of the camera is that of a positive lens forming a real image, as shown in Fig. 10A. Sharp images of distant or nearby objects are formed on a photographic film or plate, which is later developed and printed to obtain the final picture. Where the scene to be taken involves stationary objects, the cheapest of camera lenses may, if it is stopped down almost to a pinhole and a time exposure is used, yield photographs of excellent definition. If, however, the subjects are moving relative to the camera

(and this includes the case where the camera is held in the hand), extremely short exposure times are often imperative and lenses of large aperture become a necessity. The most important feature of a good camera, therefore, is that it be equipped with a lens of high relative aperture capable of covering as large an angular field as possible. Because a lens of large aperture is subject to many aberrations, designers of photographic objectives have resorted to the compromises as regards correction that best suit their particular needs. It is the intention here, therefore, to discuss briefly some of these purposes and compromises in connection with a few of the hundreds of well-known makes of photographic objective.

10.2. Speed of Lenses. It was shown in Sec. 7.15 that the total amount of light reaching the image per unit area is given by the product

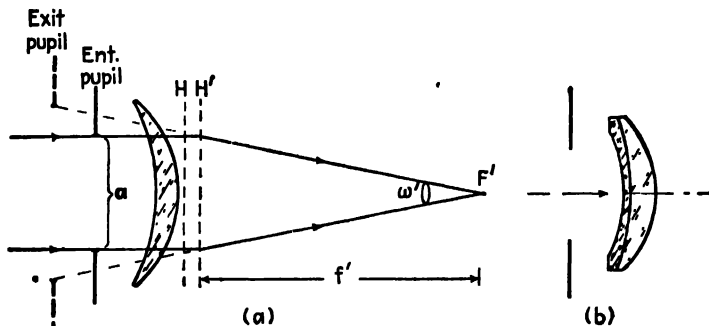


FIG. 10B. (a) Geometry for determining the speed of a lens. (b) Achromatic meniscus lens with front stop.

of the brightness B of the source and the solid angle ω' of the bundle of rays converging toward any point on the image. The latter may be computed as the area of the entrance pupil divided by the square of the focal length f . This will be clear from Fig. 10B(a), which shows the lens and stop of Fig. 10A illuminated by a parallel bundle. The solid angle ω' is that subtended at the image point by the exit pupil, but as will be seen, this is equal to that which would be subtended by the entrance pupil if it were placed at the secondary principal plane H' . The ratio of the focal length of any lens to the linear diameter a of its entrance pupil is called its *focal ratio*, or *f-value*, which is therefore defined as

$$f\text{-value} = \frac{f}{a} \quad (10a)$$

Thus a lens which has a focal length of 10 cm and a linear aperture of 2 cm is said to have an *f-value* of 5, or as it is usually stated, the lens is an $f/5$ lens.

The rapidity with which the photographic image is built up depends on the illumination E of the image, which therefore determines the *speed* of the lens. The speed is inversely proportional to the square of the f -value, since by Eq. 7*a*,

$$E = B\omega' \cong B \frac{\pi(a/2)^2}{f^2} = \text{const.} \times \frac{a^2}{f^2} = \frac{\text{const.}}{(f\text{-value})^2} \quad (10b)$$

assuming an object of a given brightness.

In order to take pictures of faintly illuminated subjects, or of ones which are in rapid motion and require a very short exposure, a lens of small f -value is required. Thus an $f/2$ lens is "faster" than an $f/4.5$ lens (or than an $f/2$ lens stopped down to $f/4.5$) in the ratio $(4.5/2)^2 = 5.06$. A lens of such large relative aperture is difficult to design, as we shall see.

10.3. Meniscus Lenses. Many of the cheapest cameras employ a single positive meniscus lens with a fixed stop such as was shown in Fig. 10*A*. Developed in about 1812 and called a *landscape lens*, this simple optical device exhibits considerable spherical aberration, thereby limiting its useful aperture to about $f/11$. Off the lens axis, the astigmatism limits the field to about 40° . The proper location of the stop results in a flat field, but with only a single lens there is always considerable chromatic aberration.

By using a cemented doublet as shown in Fig. 10*B(b)*, lateral chromatism can be corrected. Instead of correcting for the C and F lines of the spectrum, however, the combination is usually corrected for the yellow D line, near the peak sensitivity of the eye, and the blue G line, near the peak sensitivity of many photographic emulsions. Called "DG achromatism," this type of correction produces the best photographic definition at the sharpest visual focus. In some designs the lens and stop are turned around as in the arrangement of Fig. 9*U(b)*.

10.4. Symmetrical Lenses. Symmetrical lenses consist of two identical sets of thick lenses with a stop midway between them; a number of these are illustrated in Fig. 10*C*. In general, each half of the lens is corrected for lateral chromatic aberration, and by putting them together, curvature of field is eliminated, as was explained in Sec. 9.8. In the rapid rectilinear lens, flattening of the field was made possible only by the introduction of considerable astigmatism, while spherical aberration limited the aperture to about $f/8$. By introducing three different glasses, as in the Goertz "Dagor," each half of the lens could be corrected for lateral color, astigmatism, and spherical aberration. When combined they are corrected for coma, lateral color, curvature, and distortion.

Zeiss calls this lens a "Triple Protar," while Goertz calls it the "DAGor," signifying Double Anastigmat Goertz. The "Speed Panchro" lens developed by Taylor, Taylor, and Hobson in 1920 is noteworthy because of its fine central definition combined with the high speed of $f/2$ and even $f/1.5$. The "Ross" lens is but one of a number of special "wide-angle" lenses, particularly useful in aerial photography. Additional characteristics of symmetrical lenses are (1) the large number of lenses employed, and (2) the rather deep curves, which are expensive to produce.

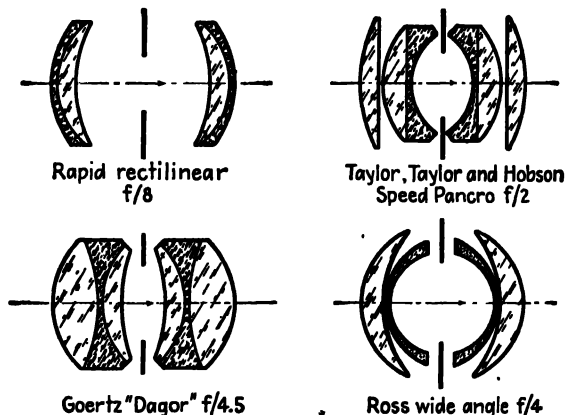


FIG. 10C. Symmetrical camera lenses.



FIG. 10D. Unsymmetrical camera lenses (triple anastigmats).

The greater the number of free glass surfaces in a lens, the greater is the amount of light lost by reflection. The f -value alone, therefore, is not the sole factor in the relative speeds of objectives. The development in recent years of lens coatings that practically eliminate reflection at normal incidence has offered greater freedom in the use of more elements in the design of camera lenses (see Sec. 14.5).

10.5. Triplet Anastigmats. A great step forward in photographic lens design was made in 1893 when H. D. Taylor of Cooke and Sons developed the "Cooke Triplet" (Fig. 10D). The fundamental principles involved in this system follow from the fact that (1) the power which a given lens contributes to a system of lenses is proportional to the height at which marginal rays pass through the lens, whereas (2) the contribution each

lens makes to field curvature is proportional to the power of the lens regardless of the distance of the rays from the axis. Hence astigmatism and curvature of field can be eliminated by making the power of the central flint element equal to the sum of the powers of the crown elements. By spacing the negative lens between the two positive lenses, the marginal rays can be made to pass through the negative lens so close to the axis that the system has an appreciable positive power. A proper selection of dispersions and radii enables additional corrections to be made for color and spherical aberration. The "Tessar," one of the best known modern photographic objectives, was developed by Zeiss in 1902. Made

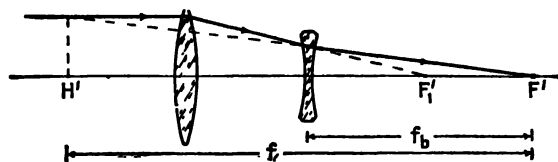


FIG. 10E. Principle of the telephoto lens.

in many forms to meet various requirements, the system has a general structure similar to that of a Cooke Triplet in which the rear crown lens is replaced by a doublet. The Leitz "Hector," working at $f/2$, is also of the Cooke Triplet type, but each element is replaced by a compound lens. This very fast lens is excellent in a motion-picture camera.

10.6. Telephoto Lenses. Since the image size for a distant object is directly proportional to the focal length of the lens, a telephoto lens which is designed to give a large image is a special type of objective with a longer effective focal length than that normally used with the same camera. Because this would require a greater extension of the bellows than most cameras will permit, the principle of a single highly corrected thick lens is modified as follows: As is shown in Fig. 10E by the refraction of an incident parallel ray, with two such lenses considerably separated the principal point H' can be placed well in front of the first lens, thereby giving a long focal length $H'F'$ with a short lens-to-focal-plane distance. The latter distance, or the "back focal length" as it is usually called, is measured from the rear lens to the focal plane, as shown.

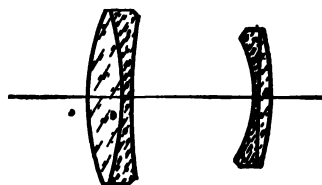


FIG. 10F. "Cooke Telephoto" of Taylor, Taylor, and Hobson.

Although the focal lengths of older types of telephoto lenses could be varied by changing the distance between the front and rear elements, they now almost always are made with a fixed focal length. Flexibility is then obtained by having a set of lenses. This has become necessary

through the desire for lenses of greater speed and better correction of the aberrations. A "Cooke Telephoto" as produced by Taylor, Taylor, and Hobson is shown in Fig. 10F.

10.7. Variable-focus Lenses. A variable-focus lens is frequently used in the motion-picture and television industries to produce what are called "zoom" shots. The desired effect is that of moving the camera away from or closer to the scene without actually transporting it. This

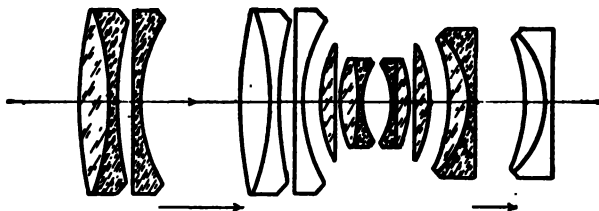


FIG. 10G. The "Zoomar," a variable-focus lens made by Taylor, Taylor, and Hobson.

is accomplished by continuously decreasing or increasing the focal length, thereby casting progressively smaller or larger images on the film or screen. There are three important factors that must be considered in the design of a variable-focus camera: first, the correction of lens aberrations; second, continual sharp focusing of the image on the film; and third, a constant f -value. All these are to be maintained automatically by the turning of a single crank or the moving of a single lever.

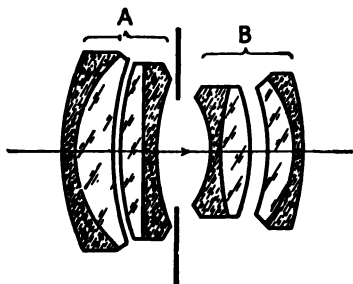


FIG. 10H. Convertible lens. Each half may be used separately, or the two of them used together.

up of two component parts separated by a stop as is shown in Fig. 10H. Elements A and B are separately corrected and may be used separately or together. Approximating a symmetrical doublet, the lens is usually made so that the front element A has a little longer focal length than element B, thus providing three different focal lengths.

10.9. The Schmidt Optical System. The Schmidt optical system combines a concave spherical mirror with an aspherical lens as shown

A variable-focus lens of high quality developed by Taylor, Taylor, and Hobson is shown in Fig. 10G. The shaded elements are movable within the ranges shown. By moving two elements, this particular combination maintains constant f -value, and the image distance as measured from the fixed elements remains constant.

10.8. Convertible Lenses. Flexibility may be added to a camera by using a convertible lens. Such a lens is usually made

in Fig. 10I. The purpose of the lens is to refract incoming parallel rays in such directions that after reflection from the spherical mirror they all come to a focus at the same axial point F . This "corrector plate," therefore, eliminates the spherical aberration of the mirror. With the lens located at the center of curvature of the mirror, parallel rays entering the system at large angles with the axis are brought to a relatively good focus at F' . The focal surface of such a system is spherical, with its center of curvature at C .

Such an optical system has several remarkable and useful properties. First as a camera, with a small film at the center or with a larger film curved to fit the focal surface, it has the very high speed of $f/0.5$. Because of this phenomenal speed, Schmidt systems are used by astronomers to obtain photographs of faint stars. They are used for similar reasons in television

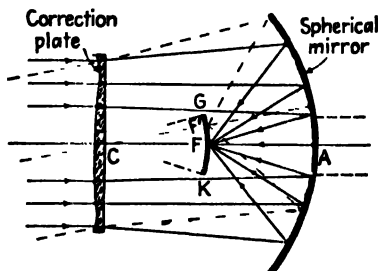


FIG. 10I. Optical arrangement for a Schmidt system.

receivers to project small images from an oscilloscope tube onto a relatively large screen. In this case the convex oscilloscope screen is curved to the focal surface GK , so that the light from the image screen is reflected by the mirror and passes through the corrector lens to the observing screen.

If a convex silvered mirror is located at GK , rays from any distant source will on entering the system form a point image on GK , and after reflection will again emerge as a parallel bundle in the exact direction of the source. When used in this manner the device is called an *auto-collimator*. If the focal surface is coated with fluorescent paint, ultraviolet light from a distant invisible source will form a bright spot at some point on GK , and the visible light emitted from this spot will emerge only in the direction of the source. If a hole is made in the center of the large mirror, an eyepiece may be inserted in the rear to view the fluorescent screen and any ultraviolet source may be seen as a visible source. As such, the device becomes a fast, wide-angled, ultraviolet telescope.

10.10. Magnifiers. The magnifier is a positive lens whose function it is to increase the size of the retinal image over and above that which is formed with the unaided eye. The apparent size of any object as seen with the unaided eye depends on the angle subtended by the object (see Fig. 10J). As the object is brought closer to the eye, from A to B to C in the diagram, accommodation permits the eye lens to change its power and to form a larger and larger retinal image. There is a limit to how close an object may come to the eye if the latter is still to have sufficient accommodation to produce a sharp image. Although the nearest

point varies widely with various individuals, 25 cm is taken to be the standard *near point*, or as it is sometimes called, the *distance of most distinct vision*. At this distance, indicated in Fig. 10K(a), the angle subtended by object or image will be called θ .

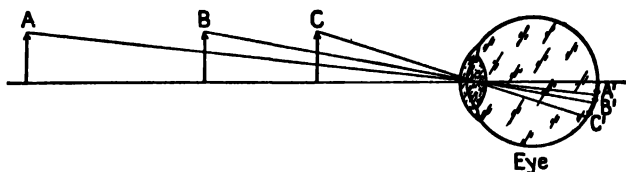


FIG. 10J. The angle subtended by the object determines the size of the retinal image.

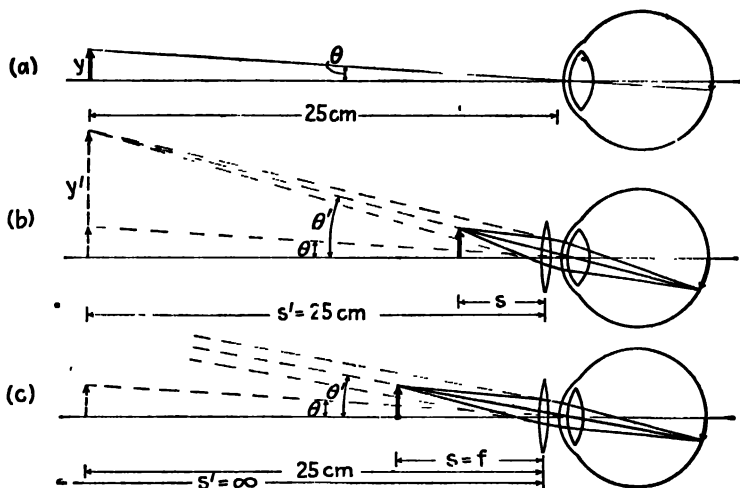


FIG. 10K. Illustrating the angle subtended (a) by an object at the near point to the naked eye, (b) by the virtual image of an object inside the focal point, (c) by the virtual image of an object at the focal point.

If a positive lens is now placed before the eye in the same position, as in diagram (b), the object y may be brought much closer to the eye and an image subtending a larger angle θ' will be formed on the retina. What the positive lens has done is to form a virtual image y' of the object y and the eye is able to focus upon this virtual image. Any lens used in this manner is called a *magnifier* or *simple microscope*. If the object y is located at F , the focal point of the magnifier, the virtual image y' will be located at infinity and the eye will be accommodated for distant vision as is illustrated in Fig. 10K(c). If the object is properly located a short distance inside of F as in diagram (b), the virtual image may be formed at the distance of most distinct vision.

The angular magnification M is defined as the ratio of the angle θ' subtended by the image to the angle θ subtended by the object.

$$M = \frac{\theta'}{\theta} \quad (10c)$$

From diagram (b), the object distance s is obtained by the regular thin-lens formula as

$$\frac{1}{s} + \frac{1}{-25} = \frac{1}{f} \quad \text{or} \quad \frac{1}{s} = \frac{25 + f}{25f}$$

From the right triangles, the angles θ and θ' are given by

$$\tan \theta = \frac{y}{25} \quad \text{and} \quad \tan \theta' = \frac{y}{s} = y \frac{25 + f}{25f}$$

For small angles the tangents can be replaced by the angles themselves to give approximate relations

$$\theta = \frac{y}{25} \quad \text{and} \quad \theta' = y \frac{25 + f}{25f}$$

giving for the magnification, from Eq. (10c),

$$M = \frac{\theta'}{\theta} = \frac{25}{f} + 1 \quad (10d)$$

In diagram (c) the object distance s is equal to the focal length, and the small angles θ and θ' are given by

$$\theta = \frac{y}{25} \quad \text{and} \quad \theta' = \frac{y}{f}$$

giving for the magnification

$$M = \frac{\theta'}{\theta} = \frac{25}{f} \quad (10e)$$

The angular magnification is therefore larger if the image is formed at the distance of most distinct vision. For example, let the focal length of a magnifier be 1 in. or 2.5 cm. For these two extreme cases, Eqs. 10d and 10e give

$$M = \frac{25}{2.5} + 1 = 11\times \quad \text{and} \quad M = \frac{25}{2.5} = 10\times$$

Because magnifiers usually have short focal lengths and therefore give approximately the same magnifying power for object distances between 25 cm and infinity, the simpler expression $25/f$, is commonly used in labeling the power of magnifiers. Hence a magnifier with a focal length

of 2.5 cm will be marked 10 \times and another with a focal length of 5 cm will be marked 5 \times , etc.

10.11. Types of Magnifiers. Several common forms of magnifiers are shown in Fig. 10L. The first, an ordinary double-convex lens, is the simplest magnifier and is commonly used as a reading glass, pocket magnifier, or watchmaker's loupe. The second is composed of two identical plano-convex lenses each mounted at the focal point of the other. As shown by Eq. 9z this spacing corrects for lateral chromatic aberration but requires the object to be located at one of the lens faces.

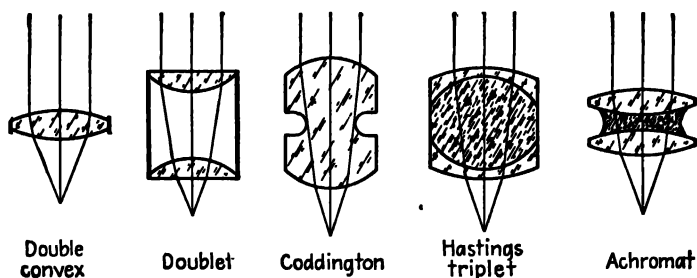


FIG. 10L. Common types of magnifiers.

To overcome this difficulty, color correction is sacrificed to some extent by placing the lenses slightly closer together, but even then the working distance is extremely short.

The third magnifier, cut from a sphere of solid glass, is commonly credited to Coddington but was originally made by Sir David Brewster. It too has a relatively short working distance, as can be seen by the marginal rays, but the image quality is remarkably good due in part to the central groove acting as a stop. Some of the best magnifiers of today are cemented triplets, such as are shown in the last two diagrams. These lenses are symmetrical to permit their use either side up. They have a relatively large working distance and are made with powers up to 20 \times .

10.12. Compound Microscopes. The compound microscope, which in general greatly exceeds the power of a simple magnifier, was invented by Galileo in 1610. In its simplest form, the modern optical microscope consists of two lenses, one of very short focus called the *objective*, and the other of somewhat longer focus called the *ocular* or *eyepiece*. While both these lenses actually contain several elements to reduce aberrations, their principal function is illustrated by single lenses in Fig. 10M. The object (1) is located just outside the focal point of the objective so there is formed a real magnified image at (2). This image becomes the object for the second lens, the eyepiece. Functioning as a magnifier, the eye-

piece forms a large virtual image at (3). This image becomes the object for the eye itself, which forms the final real image on the retina at (4).

Since the function of the objective is to form the magnified image that is observed through the eyepiece, the over-all magnification of the instrument becomes the product of the linear magnification m_1 of the objective and the angular magnification M_2 of the eyepiece. By Eqs. 3i and 10e, these are given separately by

$$m_1 = \frac{x'}{f_1} \quad \text{and} \quad M_2 = \frac{25}{f_2}$$

The over-all magnification is, therefore,

$$M = -\frac{x'}{f_1} \cdot \frac{25}{f_2} \quad (10f)$$

It is customary among manufacturers to label objectives and eyepieces according to their separate magnifications m_1 and M_2 .

10.13. Microscope Objectives. high-quality microscope is usually equipped with a turret nose carrying three objectives, each of a different magnifying power. By turning the turret, any one of the three objectives may be rotated into proper alignment with the eyepiece. Diagrams of three typical objectives are shown in Fig. 10N. The first, composed of two cemented achromats, is corrected for spherical aberration and

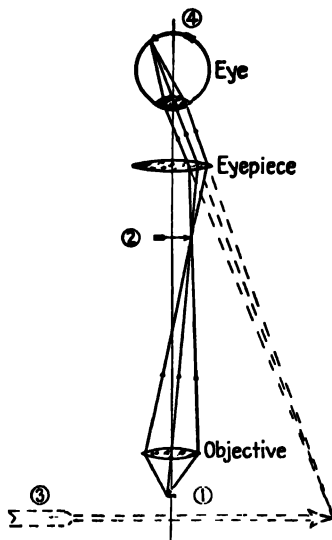


FIG. 10M. Principle of the microscope shown with the eyepiece adjusted to give the image at the distance of most distinct vision.

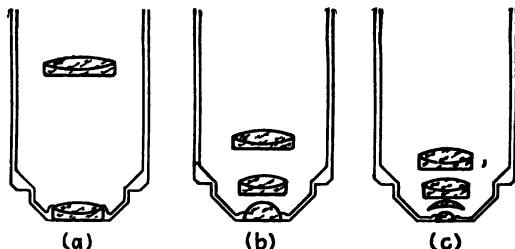


FIG. 10N. Microscope objectives: (a) low power, (b) medium power, (c) high power oil-immersion.

coma and has a focal length of 1.6 cm, a magnification of 10 \times , and a working distance of 0.7 cm. The second is also an achromatic objective with a focal length of 0.4 cm, a magnification of 40 \times , and a working

distance of 0.6 cm. The third is an oil-immersion type of objective with a focal length of 0.16 cm, a magnification of 100, and a working distance of only 0.035 cm. Great care must be exercised in using this last type of lens to prevent scratching of the hemispherical bottom lens. Although oil immersion makes the two lowest lenses aplanatic (see Fig. 9N), lateral chromatic aberration is present. The latter is corrected by the use of a compensating ocular, as will be explained in Sec. 10.20.

10.14. Astronomical Telescopes. Historically the first telescope was probably constructed in Holland in 1608 by an obscure spectacle-lens grinder, Hans Lippershey. A few months later Galileo, upon hearing that objects at a distance could be made to appear close at hand by means of two lenses, designed and made with his own hands the first authentic telescope. The elements of this telescope are still in existence

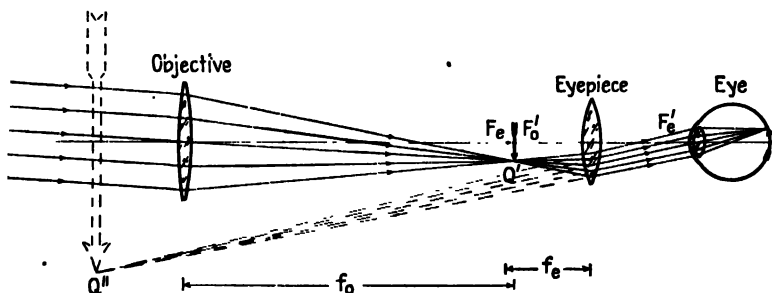


FIG. 100. Principle of the astronomical telescope shown with the eyepiece adjusted to give the image at the distance of most distinct vision.

and may be seen on exhibit in Florence. The principle of the astronomical telescopes of today is the same as that of these early devices. A diagram of an elementary telescope is shown in Fig. 100. Rays from one point of the distant object are shown entering a long-focus objective lens as a parallel beam. These rays are brought to a focus and form a point image at Q' . Assuming the distant object to be an upright arrow, this image is real and inverted as shown. The eyepiece has the same function in the telescope that it has in a microscope, namely, that of a magnifier. If the eyepiece is moved to a position where this real image lies just inside its primary focal plane F_e , a magnified virtual image at Q'' may be seen by the eye at the near point, 25 cm. Normally, however, the real image is made to coincide with the focal points of both lenses, with the result that the image rays leave the eyepiece as a parallel bundle and the virtual image is at infinity. The final image is always the one formed on the retina by rays which appear to have come from Q'' . Figure 10P is a diagram of the telescope adjusted in this manner.

In all astronomical telescopes the objective lens is the aperture stop. It is therefore the entrance pupil, and its image as formed by all the lenses to its right (here, only the eyepiece) is the exit pupil. These elements are shown in Fig. 10Q, which traces the path of one ray incident parallel to the axis and of a chief ray from a distant off-axis object point. The distance from the eye lens to the exit pupil is called the *eye relief* and should normally be about 8 mm

The magnifying power of a telescope is defined as the ratio between the angle subtended at the eye by the final image Q'' and the angle subtended at the eye by the object itself. The object, not shown in Fig.

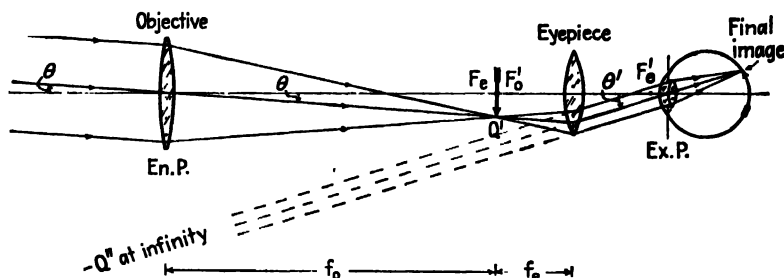


FIG. 10P. Principle of the astronomical telescope shown with the eyepiece adjusted to give the image at infinity.

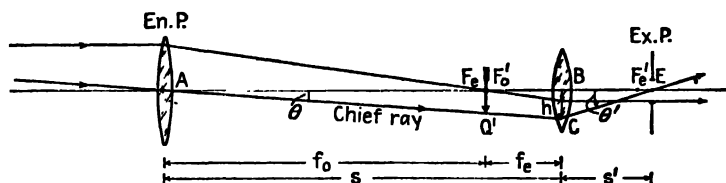


FIG. 10Q. Entrance and exit pupils of an astronomical telescope.

10Q, subtends an angle θ at the objective and would subtend approximately the same angle to the unaided eye. The angle subtended at the eye by the final image is the θ' . By definition,

$$M = \frac{\theta'}{\theta} \quad (10g)$$

From the right triangles ABC and EBC ,

$$\tan \theta = \frac{h}{s} \quad \text{and} \quad \tan \theta' = \frac{h}{s'} \quad (10h)$$

Applying the general lens formula $1/s + 1/s' = 1/f$,

$$\frac{1}{s'} - \frac{f_o}{f_e(f_o + f_e)} \quad (10i)$$

which, substituted in Eq. 10h, gives

$$\tan \theta = -\frac{h}{(f_o + f_e)} \quad \text{and} \quad \tan \theta' = \frac{hf_o}{f_e(f_o + f_e)}$$

For small angles, $\tan \theta \cong \theta$ and $\tan \theta' \cong \theta'$. Substituting them in Eq. 10g, we obtain

$$M = \frac{\theta'}{\theta} = -\frac{f_o}{f_e} \quad (10j)$$

Hence the magnifying power of a telescope is just the ratio of the focal lengths of objective and eyepiece respectively, the minus sign signifying an inverted image.

If D and d represent the diameters of the objective and exit pupil respectively, the marginal ray passing through F'_o and F_e in Fig. 10Q forms two similar right triangles, from which the following proportion is obtained

$$\frac{f_o}{f_e} = \frac{D}{d}$$

giving, as an alternative equation for the angular magnification,

$$M = \frac{D}{d} \quad (10k)$$

A useful method of determining the magnification of a telescope is, therefore, to measure the ratio of the diameters of the objective lens and of the exit pupil. The latter is readily found by focusing the telescope for infinity and then turning it toward the sky. A thin sheet of white paper held behind the eyepiece and moved back and forth will locate a sharply defined disk of light. This, the exit pupil, is commonly called the *Ramsden circle*. Its size, relative to that of the pupil of the eye, is of great importance in determining the brightness of the image and the resolving power of the instrument (see Secs. 7.15 and 15.9).

Another method of measuring the magnification of a telescope is to sight through the telescope with one eye, observing at the same time the distant object directly with the other eye. With a little practice, the image seen in the telescope can be made to overlap the smaller direct image, thereby affording a straightforward comparison of the relative heights of image and object. The object field of the astronomical telescope is determined by the angle subtended at the center of the objective by the eyepiece aperture. In other words, the eyepiece is the field stop of the system. In Fig. 10Q the angle θ is the half-field angle (Sec. 7.8).

10.15. Galilean Telescope. The first telescope made by Galileo used a negative lens as an eyepiece, as shown in Fig. 10R. In this type of instrument the focal points of the two lenses coincide beyond the eyepiece. Just as with the astronomical telescope, the magnification of this instrument is given by Eq. 10j. Since f_E is negative, M is always positive, and this means that the instrument gives an erect image.

If the objective is assumed to be the aperture stop, the exit pupil, which then is the image of the objective formed by the eyepiece, lies

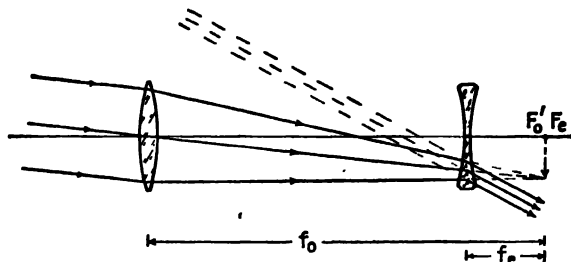


FIG. 10R. Galilean telescope. Optical system of the opera glass.

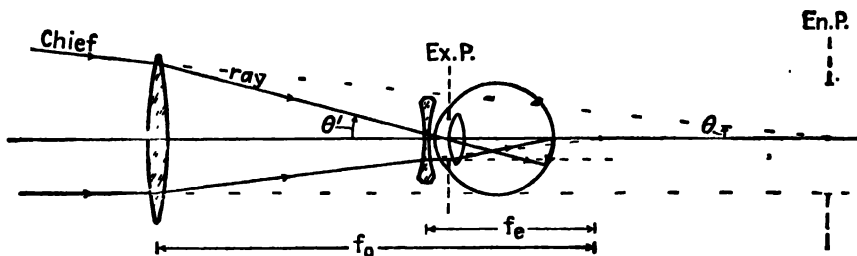


FIG. 10S. Entrance and exit pupils of a Galilean telescope.

between the eyepiece and the objective. It is not possible to place the eye at this point, and therefore the best that can be done is to place it as close to the eyepiece as possible. In this position (Fig. 10S) the pupil of the eye becomes the aperture stop and the exit pupil. Its image formed by the eyepiece and the objective is the entrance pupil. It is located behind the observer and is relatively large. A chief ray is shown approaching the margin of the objective in the direction of the entrance pupil point. After refraction it passes through the exit-pupil point. The angle θ is here the half-field angle, which indicates that the objective is the field stop of the instrument. The magnification is the ratio of the angles θ' and θ ,

$$M = \frac{\theta'}{\theta} = -\frac{f_o}{f_E} \quad (10j)$$

Since the magnification increases with f_o , the instrument is limited either to low magnification or to very small fields. The principal advantage of the Galilean telescope is its small over-all length, which makes it well suited for opera glasses. The smallness of the field of view, however, is a serious disadvantage and instruments with powers much over two are seldom made.

10.16. Oculars or Eyepieces. Although a simple magnifier of one of the types shown in Fig. 10L may be used as an eyepiece for a microscope or telescope, it is customary to design special lens combinations for each particular instrument. Such eyepieces are commonly called *oculars*. One of the most important considerations in the design of oculars is correction for lateral chromatic aberration. It is for this reason that the

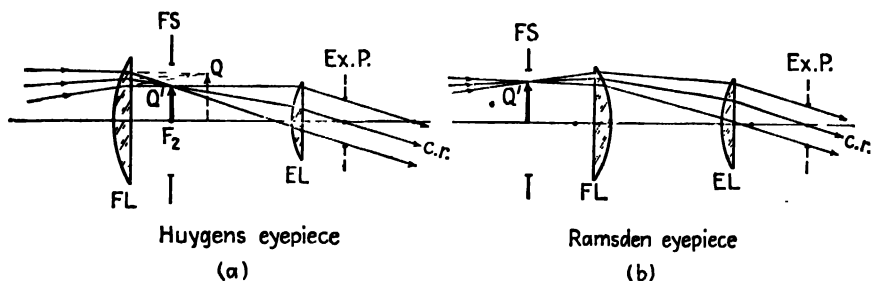


FIG. 10T. Common eyepieces used in optical instruments.

basic structure of most of them involves two lenses of the same glass and separated by a distance equal to half the sum of their focal lengths (see Eq. 9z).

The two most popular oculars based on this principle are known as the *Huygens eyepiece* and the *Ramsden eyepiece* (Fig. 10T). In both these systems the lens nearest the eye is called the *eye lens*, while the lens nearest the objective is called the *field lens*.

10.17. Huygens Eyepiece. In eyepieces of this design the two lenses are usually made of spectacle crown glass with a focal-length ratio f_1/f_2 varying from 1.5 to 3.0. As shown in Fig. 10T(a), rays from an objective to the left (and not shown) are converging to a real image point Q . The field lens refracts these rays to a real image at Q' , from which they diverge again to be refracted by the eye lens into a parallel beam. In most telescopes the objective of the instrument is the entrance pupil of the entire system. The exit pupil or *eyepoint* is, therefore, the image of the objective formed by the eyepiece and is located at the position marked "Ex. P" in the figure. Here the chief ray crosses the axis of the ocular. A field stop FS is often located at Q' , the primary focal point of the eye lens, and if cross hairs or a reticle are to be employed,

they are mounted at this point. Although the eyepiece as a whole is corrected for lateral chromatic aberration, the individual lenses are not, so that cross hairs or reticle as received through the eye lens alone will show considerable distortion and color. Huygens eyepieces with reticles are used in some microscopes, but in this case the reticle is small and is confined to the center of the field. The Huygens eyepiece shows some spherical aberration, astigmatism, and a rather large amount of longitudinal color and pincushion distortion. In general, the eye relief—*i.e.*, the distance between the eye lens and the exit pupil—is too small for comfort.

10.18. Ramsden Eyepiece. In eyepieces of this type as well, the two lenses are usually made of the same kind of glass, but here they have equal focal lengths. To correct for lateral color, their separation should be equal to the focal length. Since the first focal plane of the system coincides with the field lens, a reticle or cross hairs must be located there. Under some conditons this is considered desirable, but the fact that any dust particles on the lens surface would also be seen in sharp focus is an undesirable feature. To overcome this difficulty, the lenses are usually moved a little closer together, thus moving the focal plane forward at the sacrifice of some lateral achromatism.

The path of the rays through a Ramsden eyepiece are shown in Fig. 10T(b). The image formed by an objective (not shown) is located at the first focal point F , and it is here that a field stop FS and a reticle or cross hairs are often located. After refraction by both lenses, parallel rays emerge and reach the eye at or near the exit pupil. With regard to aberrations, the Ramsden eyepiece has more lateral color than the Huygens eyepiece but the longitudinal color is only about half as great. It has about one-fifth the spherical aberration, about half the distortion, and no coma. One important advantage over the Huygens ocular is its 50 per cent greater eye relief.

10.19. Kellner or Achromatized Ramsden Eyepiece. Because of the many desirable features of the Ramsden eyepiece, various attempts have been made to improve its chromatic defects. This aberration can be almost eliminated by making the eye lens a cemented doublet (Fig. 10U). Such eyepieces are commonly used in prism binoculars, because the slight amount of lateral color is removed and spherical aberration is reduced through the aberration characteristics of the Porro prisms (Sec. 2.2).

10.20. Special Eyepieces. The orthoscopic eyepiece shown in the middle diagram of Fig. 10U is characterized by its wide field and high magnification. It is usually employed in high-power telescopes and rangefinders. Its name is derived from the freedom from distortion character-

izing the system. The symmetrical eyepiece shown at the right in Fig. 10U has a larger aperture than a Kellner of the same focal length. This results in a wider field as well as a long eye relief, hence its frequent use in various types of telescopic gun sights. The danger of having a short eye relief with a recoiling gun should be obvious.

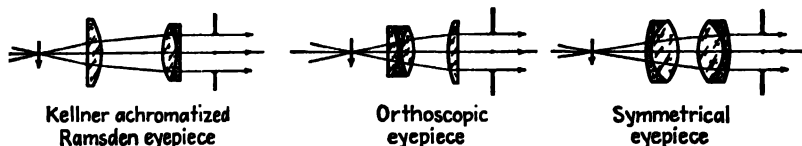


FIG. 10U. Three types of achromatized eyepieces.

Since lateral chromatic aberration, as well as the other aberrations of an eyepiece, is affected by altering the separation of the two elements, some oculars are provided with means for making this distance adjustable. Some microscopes come equipped with a set of such compensating eyepieces, thereby permitting the undercorrection of lateral color in any objective to be neutralized by an overcorrection of the eyepiece.

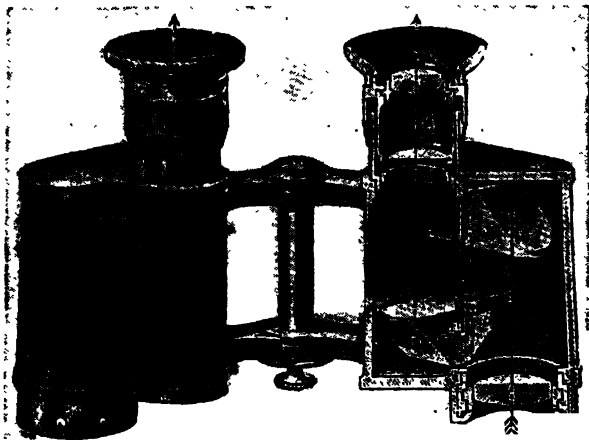


FIG. 10V. Diagram of prism binoculars showing the lenses and total reflecting Porro prisms.

10.21. Prism Binoculars. Prism binoculars are in reality a pair of identical telescopes mounted side by side, one for each of the two eyes. Such an instrument is shown in Fig. 10V with part of the case cut away to show the optical parts. The objectives are cemented achromatic pairs, while the oculars are Kellner or achromatized Ramsden eyepieces. The dotted lines show the path of an axial ray through one pair of Porro prisms. The first prism reinverts the image and the second turns it left

for right, thereby finally giving an erect image. The doubling back of the light rays has the further advantage of enabling longer focus objectives to be used in short tubes, with consequent higher magnification.

There are four general features that go to make up good binoculars: (1) magnification, (2) field of view, (3) light-gathering power, and (4) size and weight. For hand-held use, binoculars with five-, six-, seven-, or eightfold magnification are most generally used. Glasses with powers above 8 are desirable, but require a rigid mount to hold them steady. For powers less than 4, lens aberrations usually offset the magnification, and the average person can usually see better with the unaided eyes.

The field of view is determined by the eyepiece aperture and should be as large as is practicable. For seven-power binoculars a 6° object field is considered large, since in the eyepiece the same field is spread over an angle of $7 \times 6^\circ$, or 42° .

The diameter of the objective lenses determines the light-gathering power. Large diameters are important only at night when there is little light available. Binoculars with the specification 6×30 have a magnification of 6 and objective lenses with an effective diameter of 30 mm. The specification 7×50 means a magnification of 7 and

objectives 50 mm in diameter. Although glasses with the latter specifications are excellent for day or night use, they are considerably larger and more cumbersome than the daytime glasses specified as 8×30 or 8×30 . For general civilian use, the latter two are much the most useful.

10.22. Reflecting Telescopes. Most of the large astronomical telescopes in the world today employ concave paraboloidal mirrors instead of achromatic glass objectives. There are two advantages to this: (1) a concave mirror does not exhibit chromatic aberration, thus permitting a very high relative aperture, usually about $f/5$; and (2) greater stability of the telescope is attained by having the largest and heaviest optical piece at the bottom of the instrument. A diagram of the optical features of the great 200-in. reflecting telescope on Mt. Palomar is shown in Fig. 10W. With the large mirror alone, parallel rays entering the telescope tube would be brought to a focus at F . Before reaching F , however, the converging bundle is intercepted by a convex mirror of such curvature that the reflected bundle travels down the axis, through a hole

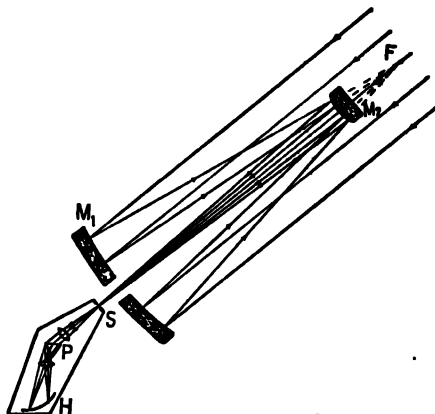


FIG. 10W. Diagram showing the principles of the 200-in. Cassegrainian telescope on Mt. Palomar.

in the center of the objective, and comes to a focus at S . A photographic plate may be located at this point, or the slit of a spectrograph as shown in the diagram.

Because the field of a reflecting telescope is free of spherical aberration and chromatic aberration, it gives extremely high definition at the field center. Off the axis, however, the oblique aberrations of coma and astigmatism increase more rapidly with obliquity than they do in a refracting telescope. The superior behavior of the refractor in these respects has, therefore, confined the use of reflecting objectives to rather large telescopes. The size limitation of a refractor is principally that of obtaining the required disks of optical glass of sufficient homogeneity to produce good lenses.

10.23. Terrestrial Telescopes. We have seen in Sec. 10.14 that an astronomical telescope composed of two positive lenses gives rise to

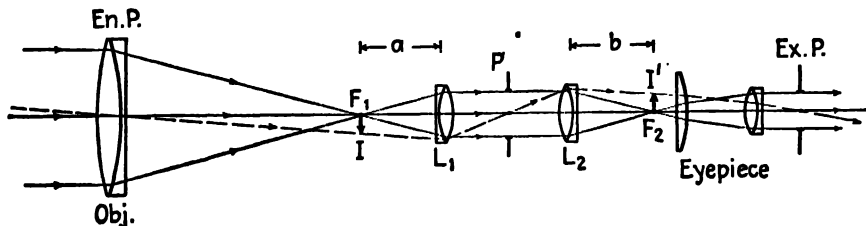


FIG. 10X. Terrestrial telescope.

inverted images and is, therefore, of little use for ground observations. Systems composed of prisms or lenses or combinations of both have been devised for erecting these images, and the final instruments are called *terrestrial telescopes*. A telescope with a two-lens erecting system is shown in Fig. 10X. Here the objective is the entrance pupil while its first image P , formed by lens L_1 , is again imaged by the lenses to its right as the exit pupil. Distant objects are first brought to a focus at F_1 , erected at F_2 , and finally imaged on the retina of the eye. Since the location of the exit pupil depends upon the location of P , the farther this image is to the right the greater is the eye relief. Because the eye relief may be too large for general use, a single positive lens is often placed at F_1 . When such a lens is used it does not alter the image positions but its aperture does determine the angular field of the instrument. By eliminating this lens, L_1 becomes the field lens.

Problems

1. A Coddington magnifier of 1 cm radius is made of crown glass of index 1.55. Applying the principle of thick lenses, find (a) its focal length, and (b) its magnifying power when the image is formed at the near point.

2. A doublet magnifier is made of two thin plano-convex lenses, each of 3 cm focal length and spaced 2 cm apart. Applying the lens formulas, find (a) its focal length, and (b) its magnifying power when the image is formed at the near point.

3. A microscope has an objective with a focal length of 0.4 cm and an ocular marked $10\times$. What is the total magnification if the objective forms its image 16 cm beyond its second focal point?

4. The objective and eyepiece of a microscope are 22 cm apart and each has a focal length of 1 cm. Treating these lenses as thin, find (a) the distance from the objective to the object viewed, (b) the linear magnification produced by the objective, and (c) the over-all magnification if the final image is seen at infinity.

5. An objective of an astronomical telescope has an aperture of 7.5 cm and a focal length of 80 cm. When it is used with an eyepiece of aperture 1 cm and focal length $f = +2$ cm, find (a) the angular magnification of a distant object, (b) the diameter of the exit pupil, (c) the angular field of the objective, (d) the angular field of the eyepiece, and (e) the eye relief.

6. The objectives of a pair of binoculars have apertures of 50 mm and focal lengths of 20 cm. The eyepieces have apertures of 1 cm and focal lengths of 2.5 cm. Find (a) the angular magnification of a distant object, (b) the diameter of the exit pupil, (c) the angular field of the objective, (d) the angular field of the eyepiece, and (e) the eye relief.

Part II
PHYSICAL OPTICS

CHAPTER 11

LIGHT WAVES

The preceding chapters were concerned with the subject of geometrical optics, the basis of which is furnished by the laws of reflection and refraction. We now turn to *physical optics*, which comprises those phenomena bearing on the nature of light. As thus defined, this field includes processes which involve the interactions of light with matter, as for example the emission and absorption of light. Many of these processes require the quantum theory for their complete explanation, but the systematic treatment of this theory lies beyond the scope of this book. A large and homogeneous class of optical phenomena can be explained by assuming that light consists of waves, and it has therefore seemed desirable to restrict the meaning of the term "physical optics" to include only the classical wave theory of light. The way in which this theory forms part of the more complete one called quantum mechanics will then be briefly described in the final chapter (Chap. 30).

As we have seen, large-scale optical effects can be explained by the use of light rays. Finer details require the wave picture which we are now to consider. Most of these details are not commonly observed in everyday life but appear when, for example, we make a close examination of the effects of passing light through narrow openings or of reflecting it from ruled surfaces. Finally, processes which occur on a still smaller scale, involving individual atoms or molecules, must be treated by quantum theory. Any case of the interaction of two or more beams of light with each other may be treated quantitatively by wave theory. As an introduction to this theory, the present chapter deals with wave motion in general and indicates at appropriate points how the various characteristics of light depend on those of the waves of which we assume it to consist.

11.1. Periodic Motion. Since the passage of a train of waves through a medium sets each particle into periodic motion, we shall first find how to give a quantitative description of this kind of motion. A periodic motion is one which repeats itself exactly in successive equal intervals of time. At the end of each interval, the particle finds itself with the same position and velocity, and the time between such occurrences is called the *period*. The simplest type of periodic motion along a straight

line is one in which the displacement y from a fixed center is given by the equation

$$y = r \sin (\omega t + \alpha) \quad (11a)$$

where t is the time and r , ω , and α are constants. This is the motion of the projection N (Fig. 11A) on the y axis of a point P moving with uniform speed in a circle of radius r . If P has the position P_0 when we start counting time ($t = 0$) and revolves counterclockwise with an angular velocity of ω rad/sec, the projection N will move up and down the y axis with a displacement y ($= ON$) given by Eq. 11a. The maximum value of the displacement is r , which is called the *amplitude* of the motion. The whole angle $(\omega t + \alpha)$ determines the position of N at any instant and is

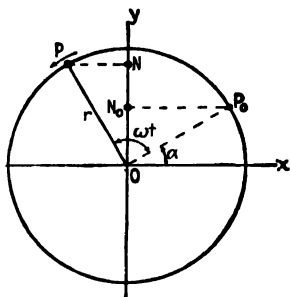


FIG. 11A. Showing how simple periodic motion may be represented by the projection on a diameter of a point P moving with uniform speed on the circumference of a circle.

called the *phase angle*, or simply the *phase*. The position N_0 at zero time is specified by the angle α , which is the initial value of the phase, or the *phase constant*. The period T is the time for a complete revolution of P . This requires a time $2\pi/\omega$ sec, since 2π is the angle swept out in a complete revolution at the rate of ω rad/sec. The period is also the time for one complete to-and-fro vibration of N along the straight line, and after the time T has elapsed the point N finds itself in the same position and with the same velocity (direction included) as it had at the beginning of this time. The reciprocal of the period is the *frequency* $\nu = 1/T$

vib/sec, because the number of vibrations performed in a time of 1 sec will be this time divided by the time T required for one vibration.

The velocity of the point N varies between zero at its extreme end-positions, when P crosses the y axis, and a maximum at the center, where P crosses the x axis. An equation for the velocity is obtained by differentiating Eq. 11a. At any instant it is given by

$$v = \frac{dy}{dt} = r\omega \cos (\omega t + \alpha) \quad (11b)$$

so that the maximum velocity is $r\omega$ or $2\pi r/T$. The acceleration a is zero at the center and a maximum at the extremes, since by differentiation of Eq. 11b,

$$\begin{aligned} a &= \frac{dv}{dt} = -r\omega^2 \sin (\omega t + \alpha) \\ &= -\omega^2 y \end{aligned} \quad (11c)$$

In the last form, the equation tells us that the acceleration is proportional to the displacement y , since we have assumed ω to be constant. The minus sign indicates that the displacement y is always opposite in direction to the acceleration. Referring to Fig. 11A, when N is above O , the acceleration is downward, and when N is below O , it is upward. According to Newton's second law, that force equals mass times acceleration, Eq. 11c means that the motion of a mass point will be given by Eq. 11a if it is acted on by a force which is proportional to the displacement and in the opposite direction. This type of motion is frequently termed simple harmonic motion or *simple periodic motion* and is physically realized in the vibrations of an elastic medium where the displacements are small and hence the forces are governed by Hooke's law.

Although simple periodic motion is evidently a very specialized type of periodic motion, it is of great importance, not only because it is frequently met with in actual waves but also because as we shall see any complex type of periodic motion can be represented as the sum of two or more such simple motions with suitable amplitudes, periods, and phase constants (Sec. 12.5). If the more complex motion is in a straight line, the component simple periodic motions from which it is made up will lie also in this line, whereas if it is confined to a plane rather than a line, we may regard it as made up of two motions (usually both complex) along two axes in this plane at right angles to each other. For example, a motion in an elliptical orbit with constant speed may be regarded as made up of two linear motions, one along the major axis and one along the minor axis of the ellipse, neither being simple periodic. If, on the other hand, the particle is attracted to the center of the ellipse with a force proportional to the distance between the particle and the center (Hooke's law), the speed in the ellipse will not be constant, but the projection along either of the axes of the ellipse or, more generally, along any other straight line through the center, will be a simple periodic motion.

Linear, circular, or elliptical vibrations are the types most frequently dealt with in the study of light waves. The vibrating source, which is necessary for the emission of any kind of waves, may for certain cases of the emission of light be thought of as an electron revolving about the nucleus of an atom. For a large orbit, the force exerted by the rest of the atom on the electron in question will vary approximately as the inverse square of the distance (Coulomb's law), and the orbit will be nearly an ellipse with the nucleus of the atom at one focus. Classically the vibrations in the light emitted in a direction perpendicular to the plane of the orbit will then have an elliptical form corresponding to that of the electron orbit, while in the light emitted in the plane of the orbit

they will have a linear form, corresponding to the motion seen when the ellipse is viewed edge-on. Figure 11B shows the orbit of an electron in an atom such as sodium.

11.2. Wave Motion. Waves of the type with which we are most familiar, *i.e.*, waves on the surface of water, are of considerable complexity. However, they may serve to illustrate an important feature present in any wave motion. If the waves are traveling in the x direction and the y direction is vertical, an instantaneous picture of the con-

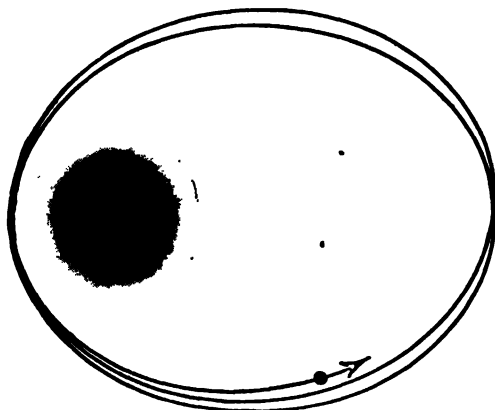


FIG. 11B. Schematic diagram of a sodium atom with its single orbital electron.

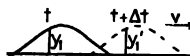


FIG. 11C. Illustrating the propagation of water waves.

tour of the waves in the x, y plane is given in Fig. 11C by the continuous curve. Let this curve be represented by an equation $y = f(x)$. If the wave contour is to move toward $+x$ with a constant velocity v , we must introduce the time t in such a way that, as t increases, a given ordinate such as y_1 will, after a time Δt has elapsed, be found at y'_1 , farther to the right by an amount $\Delta x = v \Delta t$. This is accomplished by writing the equation $y = f(x - vt)$, since we have, at the two times t and $t + \Delta t$,

$$\begin{aligned} y_1 &= f(x - vt) \\ y'_1 &= f[(x + \Delta x) - v(t + \Delta t)] \end{aligned}$$

If now we substitute $\Delta x = v \Delta t$, we find that $y'_1 = y_1$, and the above requirement is realized. The wave is in the position of the broken curve

at the instant $t + \Delta t$. The general equation for any transverse wave motion in a plane is

$$y = f(x \pm vt) \quad (11d)$$

The plus sign is to be used if the wave is to travel to the left, i.e., in the $-x$ direction.

The reader should not infer from the foregoing discussion that the particles of water are transferred to the right along with the wave. On the contrary, the only thing that moves along continuously is the contour, while each particle merely oscillates about its position of equilibrium. For water waves the motion of each particle is circular or elliptical in the x, y plane. In this case the ordinate y is merely the y component of the displacement of the particle from its equilibrium position, since the motion is not a transverse one confined to the y direction. Hence we next consider the simplest type of waves, where this complication does not arise.

11.3. Simple Periodic Waves. Suppose that the wave contour $y = f(x)$ is given by

$$y = r \sin \left(-\frac{2\pi}{\lambda} x \right) \quad (11e)$$

and that the displacements are strictly transverse. The significance of the constants r and λ may be seen from Fig. 11D, which is a plot of

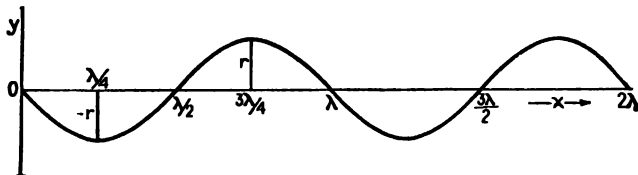


FIG. 11D. Wave contour of a simple periodic wave.

Eq. 11e. The maximum displacement is r , which then represents the *amplitude* of the wave, while λ is the distance after which the curve repeats itself and is called the *wavelength*. If now the wave is to progress toward $+x$, Eq. 11d tells us that we must substitute $(x - vt)$ for x in Eq. 11e. We thus have

$$y = r \sin \frac{-2\pi}{\lambda} (x - vt) = r \sin \frac{2\pi}{\lambda} (vt - x) \quad (11f)$$

Any particular particle is specified by one value of x , and as the wave contour moves by this point with a velocity v , the particle moves up and down with a displacement determined by the ordinate of the curve. The amplitude of its motion is evidently r and the period

$$T = \frac{\lambda}{v} \quad (11g)$$

since a complete vibration is executed as the wave travels a distance λ . The equation of the motion of one particle is obtained by giving x a particular value in Eq. 11f. For the particle with $x = 0$,

$$y = r \sin \frac{2\pi vt}{\lambda} = r \sin \left(\frac{2\pi}{T} \right) t \quad (11h)$$

This is the same as Eq. 11a with $\omega = 2\pi/T$ and $\alpha = 0$; hence the particle moves with simple periodic motion. The same is true of every other particle, and the motions are identical except for a progressive decrease in the phase constant α with increasing x . Figure 11E shows the posi-

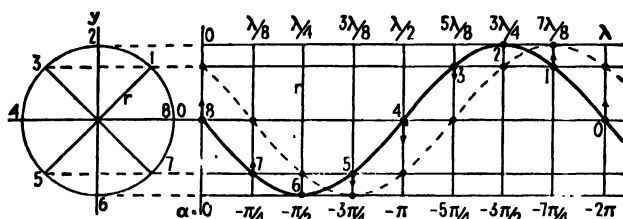


FIG. 11E. Illustrating the propagation of a simple periodic wave.

tions of particles $\frac{1}{8} \lambda$ apart at the times $t = 0$ (solid curve) and $t = \frac{1}{8} T$ (dotted curve), with the phase constants of the various points shown below each. The velocity of each particle is indicated by the attached arrow. The figure also illustrates a convenient way to plot a sine curve. We see that the wave contour has progressed $\frac{1}{8} \lambda$ toward the right in a time $\frac{1}{8} T$.

Making use of the relation $\lambda = vT$ from Eq. 11g, Eq. 11f may be written

$$y = r \sin \frac{2\pi}{vT} (vt - x) = r \sin \frac{2\pi}{T} \left(t - \frac{x}{v} \right) \quad (11i)$$

Another very convenient form, easily remembered because of its symmetry, is

$$y = r \sin 2\pi \left(\frac{t}{T} - \frac{x}{vT} \right) = r \sin 2\pi \left(\frac{t}{T} - \frac{x}{\lambda} \right) \quad (11j)$$

Here we see immediately the dependence of the phase constant α on x , since $\alpha = -\frac{2\pi}{\lambda} x$. The phase constant of a particle is not a particularly significant quantity, since its value will depend upon the choice of the

zero of time. However, the *phase difference* of two particles is independent of this choice, since we have, for two particles at x_1 and x_2 ,

$$\begin{aligned}\text{Phase difference} &= 2\pi \left(\frac{t}{T} - \frac{x_1}{\lambda} \right) - 2\pi \left(\frac{t}{T} - \frac{x_2}{\lambda} \right) \\ &= \frac{2\pi}{\lambda} (x_2 - x_1)\end{aligned}\quad (11k)$$

Since the phase decreases with increasing x , particle 1 is advanced in phase with respect to particle 2 when $x_2 > x_1$. If two trains of identical waves start in the same phase from a given point, the difference between the phase of a particle at a distance x_1 along one wave and that of another particle x_2 along the other wave will be that given by Eq. 11k. In this case we speak of the retardation, or *path difference*, $x_2 - x_1$, and we shall in the following chapters frequently have occasion to convert path differences into phase differences by the relation

$$\text{Phase difference} = \frac{2\pi}{\lambda} \cdot (\text{path difference}) \quad (11l)$$

Light which is exactly described by the above equations for simple periodic waves is said to be perfectly *monochromatic plane-polarized* light; i.e., it possesses one accurately defined wavelength, and the vibration is confined to one plane containing the direction of propagation. As we shall see later, it is impossible to produce light which fulfills these requirements strictly. One obvious reason for this is that Eq. 11j requires a train of waves infinitely long, extending from $x = +\infty$ to $x = -\infty$.

11.4. Velocity of Waves. Waves such as those we have described may be generated in a horizontal stretched rope by moving one end up and down with a simple periodic motion. The vibrating mechanism at the end of the rope then acts as a source and when the end is displaced a force is exerted on the adjacent element of the rope by virtue of its tension. This, in turn, disturbs the next element, and thus the wave is propagated along the rope. Each new element as it is disturbed lags behind the preceding one, due to its inertia, and this is the cause of the progressive change in phase along the wave. The rate at which the disturbance is propagated will evidently be greater the greater the force acting between adjacent elements. In a stretched rope this force is furnished by the tension. The rate will be smaller when the inertia of the elements is larger, i.e., for a heavier rope.

It is not difficult to derive an equation for the velocity of a wave in a flexible rope. Consider a wave of the form represented in Fig. 11F(a) traveling to the right with the velocity v . This velocity is the same as

that with which the rope would have to move to the left in order to keep the wave stationary with respect to the observer [Fig. 11F(b)]. If F is the tension in the rope, the resultant force on a short element RQ , of length l , will be f , the vector sum of the two tensions F acting on the ends of the element. This force f produces centripetal acceleration of the element as its center passes from R to Q . Representing the velocities v_R and v_Q at R and Q by vectors in Fig. 11F(b), this acceleration is determined by the vector difference Δv , such that

$$a = \frac{\Delta v}{t}$$

If m is the mass per unit length of the rope, the mass of the element is ml . By Newton's second law of motion

$$f = mla$$

FIG. 11F. Illustrating the action of forces propagating a wave along a rope.

The triangle of velocities in the figure is similar to the triangle of forces, so

$$\frac{\Delta v}{v} = \frac{f}{F}$$

whence

$$\frac{at}{v} = \frac{mla}{F}$$

But, since $t = l/v$, we obtain

$$v = \sqrt{\frac{F}{m}}$$

The tension F must, of course, be expressed in absolute units of force, i.e., dynes in the cgs system.

In an extended medium of three dimensions, waves of the type described by Eq. 11j exist only if the medium has rigidity, or resistance to shear. This property is a characteristic of solids, so that such waves may be set up in a solid by a suitable vibrating source. Any point in the solid when displaced from its equilibrium and released will move with simple periodic motion by virtue of the fact that for small displacements the forces of elasticity obey Hooke's law. Two sets of waves will be sent out, one of *longitudinal waves*, in which the vibrations are along the direction of motion of the waves, and one of *transverse waves*, of the type described above. The mathematical investigation of the velocity of these waves leads to a formula of the same type as the above equation for the waves in a rope. Instead of the tension, the measure of the restoring force now becomes a quantity E depending on the elastic constants of the medium.

The linear density of the rope is replaced by the volume density ρ of the medium. This gives

$$v = \sqrt{\frac{E}{\rho}}$$

For transverse waves, which occur only in solids, $E = n$, the rigidity modulus. For longitudinal waves in a liquid, $E = k$, k being the bulk modulus, while in an extended solid $E = k + \frac{4}{3}n$. Of the two types of waves in a solid, the longitudinal waves obviously travel the faster.

All light waves, as we shall see, are essentially *transverse*. Their velocity in vacuum, measured by methods to be described in Chap. 19, is about 3×10^{10} cm/sec and is independent of wavelength and intensity. Since the possibility of transverse waves requires a medium endowed with rigidity, the "elastic-solid theory" postulated an all-pervading medium, called the ether, for the propagation of light. Because of the large velocity of light, the ether was supposed to have a high rigidity and small density. As we shall see, this theory constitutes a useful working hypothesis but leads to some serious inconsistencies. It has now been definitely abandoned in favor of the electromagnetic theory (Chap. 20). In a transparent substance, the measured velocity of light is always less than in free space. The elastic solid theory explained this as due to either an increased density of the ether within matter or a decreased rigidity. Neither of these assumptions is in agreement with all the observed facts.

11.5. Wavelength. The wavelength of a train of waves like that in Fig. 11E is the distance between one particle and the nearest one which is in the same phase. In most important types of waves, the vibrations are not confined to a single line of particles but traverse a medium of three dimensions. In this case the motion of any one particle in Fig. 11E may be taken as representative for all of the particles which lie on a *wave front*, which we may define as the locus of immediately adjacent points vibrating in the same phase. As thus defined, this term, originally introduced in Sec. 1.8 to describe a surface normal to the rays, is now given a physical significance. The wave front may in general have any shape, but those most commonly found are plane or spherical. Plane wave fronts would be produced in a block of elastic material (Fig. 11G)

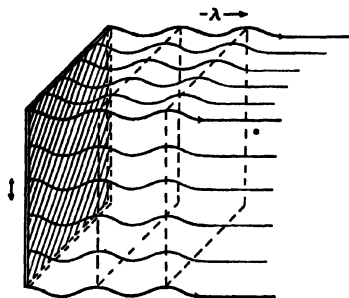


FIG. 11G. Illustrating the generation of a plane wave in an elastic solid.

by giving a periodic motion in its own plane to a board attached to one surface of the block. Here the motion of any particle in a wave front, such as one of those indicated by the dotted plane in Fig. 11G, is in all respects identical with that of any other particle in that wave front. Such *plane waves* are practically attained at a sufficient distance from a *point source* (Fig. 11H). The waves only a few wavelengths away from the source are *spherical waves*, but as we go farther away a restricted portion of the wave front becomes more and more nearly plane. For spherical waves, the motion of each particle in the wave front is the same as

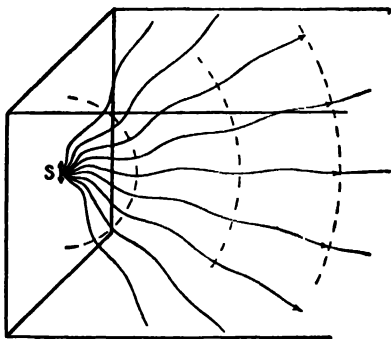


FIG. 11H. Illustrating spherical waves in an elastic solid.

that of any other particle except for a difference in the direction of vibration on different parts of the wave front, these being always tangent to the spherical surface. In either plane or spherical waves, the perpendicular distance between two adjacent wave fronts having the same phase of vibration is the wavelength.

When a train of waves passes from one medium into another, there will be a change of velocity due to the changed elasticity and density. This will entail a change in wavelength as well. In the case of light the ratio of the velocity in free space to that in a substance is known as the *index of refraction* n of that substance:

$$n = \frac{c}{v} \quad (11m)$$

where we use the symbol c for the velocity in free space (3×10^{10} cm/sec). Now when the velocity of a wave changes, either the wavelength or the frequency, or both, must change, since we have, from Eq. 11g,

$$v = \frac{\lambda}{T} = \nu\lambda \quad (11n)$$

ν being the frequency. Since there is no reason for the rate of vibration to be altered in passing through a boundary, we should expect the frequency to be unchanged, and λ to be changed when v changes. In fact, measurements show that λ becomes smaller in the medium in exactly the proportion that v decreases.

In calculating the phase difference between two particles in a train of waves from Eq. 11f, account must be taken of the fact that λ is changed

if the velocity is appreciably different from that in vacuum. We are accustomed to speaking of the wavelength of light λ as its value in air or in vacuum, but this is not the wavelength to be used in calculating path differences within the medium. If we denote by λ_m the wavelength in the medium, Eq. 11*k* should be written

$$\text{Phase difference } \delta = \frac{2\pi}{\lambda_m} \cdot (x_2 - x_1)$$

Now, by what has been said above,

$$\frac{\lambda}{\lambda_m} = \frac{c}{v} = n$$

so that

$$\delta = \frac{2\pi c}{\lambda v} (x_2 - x_1) = \frac{2\pi}{\lambda} \cdot n(x_2 - x_1) \quad (11o)$$

Therefore we may apply Eq. 11*l* as it stands for calculating phase difference from path difference in cases where light traverses a medium, provided that for the path difference we use a quantity Δ , representing the product of the geometrical path difference within the medium by the index of refraction. This quantity is the difference in *optical path* [d], as defined and used in the first chapter (Sec. 1.4). The optical path difference is defined by the equation

$$\Delta = [d_2] - [d_1] = n(x_2 - x_1) \quad (11p)$$

Physically, the optical path is the distance in vacuum containing the same number of waves as the actual geometrical path in the medium.

The wavelengths of visible light extend between about 4×10^{-5} cm for the extreme violet and 7.2×10^{-5} cm for the deep red. Just as the ear becomes insensitive to sound above a certain frequency, so the eye fails to respond to light vibrations of frequencies greater than that of the extreme violet or less than that of the extreme red. The limits, of course, depend somewhat upon the individual, and there is evidence that most persons can see an image with light of wavelength as short as 3.0×10^{-5} cm, but this is a case of fluorescence in the retina. In this case the light appears to be a bluish gray in color and is harmful to the eye. Radiation of wavelength shorter than that of the visible is termed "ultraviolet light" down to a wavelength of about 5×10^{-7} cm, and beyond this we are in the region of X rays to 6×10^{-10} cm. Shorter than these, in turn, are the γ rays from radioactive substances. On the long-wavelength side of the visible lies the infrared, which may be said

to merge into the radio waves at about 4×10^{-2} cm. Figure 11I shows the names which have been given to the various regions of the spectrum of radiation, though we know that no real lines of demarcation exist. It is not convenient to use the same units of length throughout such an enormous range. Hence radio wavelengths are expressed in meters (10^2 cm), infrared in microns ($1 \mu = 10^{-4}$ cm), visible and ultraviolet in angstrom* units ($1 \text{ A} = 10^{-8}$ cm) and X rays in angstroms, or, commonly in accurate work, in X units ($1 \text{ XU} = 10^{-11}$ cm).

It will be seen that visible light covers an almost insignificant fraction of this range. Therefore, although all these radiations are similar in nature, differing only in wavelength, the term "light" is conventionally

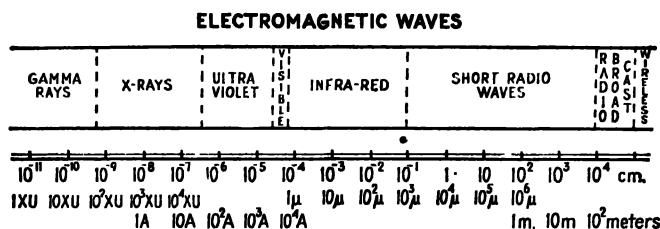


FIG. 11I. Scale of wavelengths for the range of known electromagnetic waves.

extended only to the adjacent portions of the spectrum, namely, the ultraviolet and infrared. Many of the results that we shall discuss for light are common to the whole range of radiation, but naturally there are qualitative differences in behavior between the very long and very short waves, which we shall occasionally point out. The divisions between the different types of radiation are purely formal and are roughly fixed by the fact that in the laboratory the different types are generated and detected in different ways. Thus the infrared is emitted copiously by hot bodies, and is detected by an energy-measuring instrument such as the thermopile. The shortest radio waves are generated by electric discharges between fine metallic particles immersed in oil and are detected by electrical devices. Nichols and Tear, in 1917, produced infrared waves having wavelengths up to 0.42 mm and radio waves down to 0.22 mm. The two regions may therefore be said to overlap, keeping in mind, however, that the waves themselves are of the same nature for both. The same holds true for the boundaries of all the other regions of the spectrum.

11.6. Doppler Effect. When a source of waves is in motion through a stationary medium, the wavelength is changed. The waves sent out in

* A. J. Ångström (1814-1874). Professor of physics at Uppsala, Sweden. Chiefly known for his famous atlas of the solar spectrum, which was used for many years as the standard for wavelength determinations.

the direction of motion of the source are shorter, and those in the opposite direction longer, than the waves from the source at rest. This is the Doppler* effect, which is responsible for the well-known change of pitch of sound when a source passes the observer at high speed. An equation for the new wavelength is easily obtained by a consideration of Fig. 11J(a) in which are pictured the waves sent out by a source which is moving toward the right. The circle represents a section of the spherical wave front one whole period T after it left the source at S . If the source were stationary at S , the wave would have traveled just one wavelength λ and would at this instant be represented by the dotted curve. Instead, during one period the source moves a distance vT toward the right, where v is its velocity. As a result, the wavelength is shortened by the distance

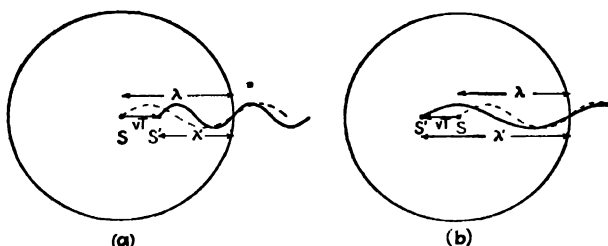


FIG. 11J. (a) Source of waves moving in the direction of wave propagation. (b) Source of waves moving opposite to the direction of wave propagation.

SS' , which equals vT , and the actual wave will be represented by the solid curve. To find the new wavelength λ' we note that

$$\lambda - \lambda' = vT$$

Calling the velocity of the waves c , we have, from Eq. 11g, $\lambda = cT$, and hence

$$\lambda' = cT - vT = T(c - v) = \lambda \left(\frac{c - v}{c} \right)$$

From Fig. 11J(b), in which the source is moving in the opposite direction, one obtains in the same way

$$\lambda' = cT + vT = \lambda \left(\frac{c + v}{c} \right)$$

* J. C. Doppler (1803–1853). Native of Salzburg, Austria. At the age of thirty-two, unable to secure a position, he was about to emigrate to America. However, at that time he was made professor of mathematics at the Realschule in Prague, and later became professor of experimental physics at the University of Vienna.

The two equations may be combined in the form

$$\lambda' = \lambda \left(1 \pm \frac{v}{c} \right) \quad (11g)$$

the positive sign referring to the case where the source is moving in the opposite direction from the waves. For the part of the wave traveling off at any angle to the direction of motion of the source, the changed wavelength will obviously lie between these two values. The velocity of the waves is not changed by the motion of the source, so that the new wavelength corresponds to a new frequency

$$\nu' = \frac{c}{\lambda'} = \frac{c}{\lambda \left(\frac{c \pm v}{c} \right)} = \frac{\nu}{1 \pm \frac{v}{c}} \quad \text{SOURCE IN MOTION} \quad (11r)$$

A change of frequency is also encountered when the source is at rest, but the observer moving toward or away from the source. In Fig. 11K

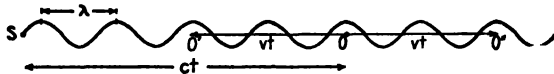


FIG. 11K. Doppler effect with the observer moving toward or away from a stationary source.

let the source S send out a train of waves toward the right, and let the observer O be stationary at a distance $SO = ct$ from the source. In the time t , he will receive the whole train of waves between S and O , and the observed frequency will be the number of waves passing him per second, or

$$\nu = \frac{ct/\lambda}{t} = \frac{c}{\lambda}$$

in agreement with Eq. 11g. If, now, the observer moves toward the source with a velocity v , he will move from O to O' , a distance vt , in the above time t . As a result, he will receive vt/λ extra waves, and will observe an increase in frequency of $(vt/\lambda)/t$, or v/λ . Hence the apparent frequency will be

$$\nu' = \frac{c}{\lambda} + \frac{v}{\lambda}$$

and, since $\lambda = c/\nu$,

$$\nu' = \nu \left(\frac{c + v}{c} \right)$$

If, instead, the observer moves from O to O'' , away from the source, he fails to receive the number v/λ of waves included in this distance, and

the frequency is now lowered to the value

$$\nu' = \nu \left(\frac{c - v}{c} \right)$$

Including both cases in one equation, we may write

$$\nu' = \nu \left(1 \pm \frac{v}{c} \right) \quad \text{OBSERVER IN MOTION} \quad (11s)$$

where the positive sign is to be used when the observer is approaching the source.

In sound and other mechanical waves, there is a definite physical difference between the two cases (source in motion or observer in motion) discussed above. In the first case the character of the waves is altered by a real change of wavelength. In the second the wavelength is unchanged, but there is an apparent change of frequency to an observer in motion. Correspondingly, the two equations 11r and 11s give different results for the same v , especially when v is not small compared to c . To compare the two equations, let us expand the right-hand member in Eq. 11r by the binomial theorem. This gives

$$\nu' = \nu \left(1 \pm \frac{v}{c} + \frac{v^2}{c^2} \pm \frac{v^3}{c^3} + \cdots \right)$$

which is the same as 1s except for the small terms v^2/c^2 , etc. For sound, c is about 1100 ft/sec, and with a velocity v of 60 m.p.h., v^2/c^2 is only 0.0064, and the terms of higher order much smaller still. This difference between Eqs. 11r and 11s, though small, is still appreciable for sound when large velocities are involved.

When we consider the Doppler effect for light, we find the above difference so small that it is entirely negligible in most cases, owing to the extremely high velocity of light. For any velocity attainable on earth even the term v/c is so small that it is very difficult to detect. The stars, however, have velocities toward or away from the earth usually ranging between about 10 and 30 km/sec, with some as high as 300 km/sec. If the spectrum of such a star is photographed alongside that of an arc or spark giving some of the same lines appearing in the stellar spectrum, the corresponding stellar lines are found to be shifted slightly toward the violet or toward the red, according to whether the star is approaching or receding from the earth. Figure 11L shows, in the center strip, the spectrum of the star μ Cassiopeiae. Above and below are the lines of iron from a laboratory source. All these iron lines appear as white lines in the star spectrum, but they are displaced toward the left,

i.e., toward shorter wavelengths. This increase of frequency shows that the star is approaching the earth, and measurement of this spectrogram gives the unusually high velocity of 115 km/sec.

The change of wavelength observed here may be regarded either as a real change of the length of the waves, due to the motion of the source, or as an apparent change due to the motion of the observer with the earth. It is immaterial in light whether we use Eq. 11r or Eq. 11s, in calculating the velocity, since, even for a star with the enormous velocity of 300 km/sec, the principal term by which these equations differ, v^2/c^2 , is only 1 part in 1 million. It should be emphasized, however, that for light there is really no physical difference between the two cases, source

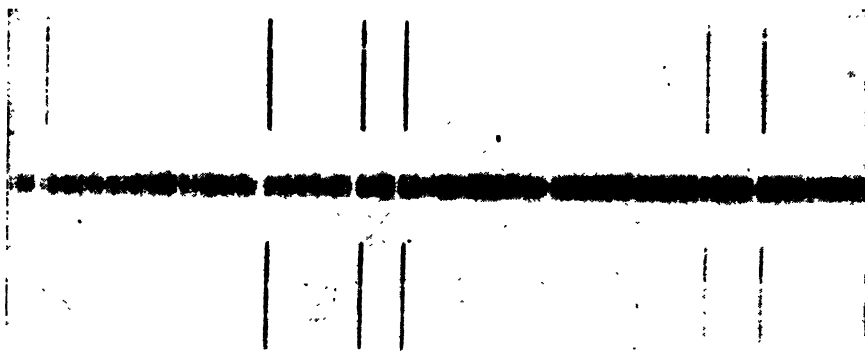


FIG. 11L. Doppler shift of spectrum lines in a star. (Spectrum taken at the Dominion Astrophysical Observatory.)

in motion and observer in motion, because this distinction rests upon the assumption that there is a medium which we may consider at rest, such as the air for sound, and relative to which we express these motions. Now, according to the theory of relativity, we are not justified in assuming such a fixed ether, having a definite location in space (Chap. 19). Any body may be assumed at rest, and if all other motions are expressed relative to it, the results will be the same regardless of which body we choose. With this postulate, the theory of relativity leads to the equation

$$\nu' = \nu \left(1 \pm \frac{v}{c} + \frac{1}{2} \frac{v^2}{c^2} \pm \frac{1}{2} \frac{v^3}{c^3} + \dots \right) \quad \text{FOR LIGHT}$$

It is interesting that the change of frequency predicted by this formula lies just halfway between those given by Eqs. 11r and 11s.

The Doppler effect for light has been applied to a number of other astronomical problems. For example, it is possible to measure the rotation of the sun by observing the spectra of the east and west edges.

The lines show a displacement corresponding to a rotational velocity of 2.1 km/sec. Some other astronomical applications are the detection of close double stars, the rotation of the planets and of Saturn's rings, motions of clouds of luminous matter on the sun's surface, and the apparent motion of distant universes (spiral nebulae). In the latter case, lines have been observed to be shifted toward the red by more than 200 Å, the corresponding velocities of recession being more than 20,000 km/sec.

In the laboratory, the Doppler shift produced by reflecting light from mirrors mounted on the rim of a wheel rotating at high speed has been detected. It also provides a method of investigating the motions of atoms in a source of light.

11.7. Amplitude. The amplitude of a set of waves is the maximum displacement of any particle from its equilibrium position. Its importance lies in the fact that it determines the amount of energy carried by

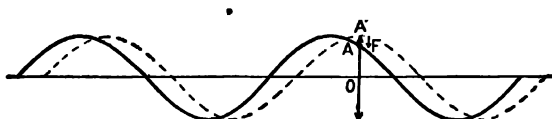


FIG. 11M. Illustrating the amplitude and periodic displacement of a particle due to a passing wave.

the waves. The amount of energy flowing per second across unit area of a surface normal to that surface is called the *intensity*, and hence we are interested in the relation between the energy of waves and their amplitude. This may be found in a very elementary manner for the case where a displaced particle is drawn back toward its equilibrium position by a force directly proportional to the displacement, *i.e.*, one that obeys Hooke's law. We have shown in Sec. 11.1 that this is the requisite condition for simple periodic motion, which is the type of motion involved in the simple periodic waves discussed above. In Fig. 11M, let a train of such waves be traveling to the right, and suppose that they have set the particle *A* into simple periodic motion along the line indicated by the double-pointed arrow. At the instant when the wave occupies the position indicated by the continuous curve, let the displacement *AO* from the center be *y*; then, by Hooke's law, the return force

$$F = -ky$$

where *k* is a constant. Because of its inertia, the particle will continue to move outward against this force *F*, which is the resultant of the forces exerted on it by adjacent particles. When the wave has reached the

position of the broken curve, the particle will have reached A' . The displacement now has its maximum value $A'O = r$, which we call the amplitude. In this position the particle is instantaneously at rest, acted on by a force $-kr$, and hence possesses potential energy. The kinetic energy is zero at this instant, and as the particle begins to move down, it gains kinetic energy at the expense of potential energy only, so that the total energy remains constant and equal to the maximum value of the potential energy. Therefore, to find the total energy of the vibration, we need only evaluate the potential energy at A' . This equals the work done in overcoming the force F between O and A' , and since F is variable, we may find the work from the integral

$$\int_0^r -F \, dy = \int_0^r ky \, dy = \frac{1}{2}kr^2$$

The vibrational energy imparted to any one particle is therefore proportional to the square of the amplitude. The amount of energy per unit length, or per unit volume, of the wave will also be proportional to r^2 , since this merely increases the energy by a factor equal to the number of particles in the unit. The intensity is also proportional to the square of the amplitude, since it is the energy contained in a volume of unit cross section and of length v (where v is the velocity of the waves), and this is the volume which will cross unit area per second.

When a set of spherical waves diverges from a vibrating point in a perfectly transparent medium, as in Fig. 11H, the intensity, as defined above, decreases in a given direction inversely as the square of the distance from the source. This inverse-square law of intensities holds quite generally for spherical waves emitted by a source of dimensions small compared to the distances at which the intensity is measured. In practice, it is sufficient that the distance be greater than ten times the lateral dimensions of the source, since at this distance the error made by assuming the inverse-square law is already less than 0.1 per cent. The inverse-square law is easily proved from the fact that, when spherical waves diverge from a point source, the same amount of energy must pass per second through any sphere drawn with the source as its center. Since the areas of such spheres increase directly as the squares of their radii, the energy falling on unit area per second, or the intensity at any distance d is inversely proportional to d^2 , so that, if I_1 , and I_2 are the intensities at two distances d_1 and d_2 , we have, as an expression of the inverse-square law,

$$\frac{I_1}{I_2} = \frac{d_2^2}{d_1^2} \quad (11f)$$

Stated in terms of the amplitude, the law requires that the amplitude change inversely as the first power of the distance, since the intensity is measured by the square of the amplitude. Equation 11j may therefore be written, for spherical waves,

$$y = \frac{r}{d} \sin 2\pi \left(\frac{t}{T} - \frac{d}{\lambda} \right) \quad (11u)$$

r now being the amplitude at unit distance from the source.

If the medium is not perfectly transparent, there will be an additional decrease in amplitude due to absorption and scattering (Chap. 22). No natural substance is perfectly transparent, since some loss of intensity can always be detected by using sufficient thicknesses, and hence the inverse-square law is strictly applicable only in a vacuum. In a material medium, the amplitude changes because of absorption as the waves proceed through the medium. Bouguer's* law of absorption states that layers

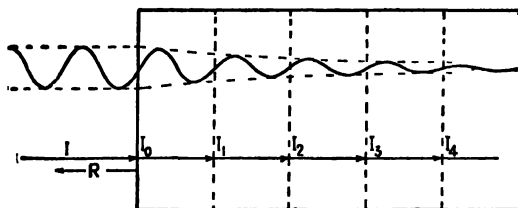


FIG. 11N. Illustrating the diminishing amplitude of a wave passing through an absorbing medium.

of equal thickness absorb equal fractions of the intensity incident upon them, whatever this intensity may be. Thus, in Fig. 11N, let I be the intensity of the waves incident on the front surface of the medium. Of this, an amount R will be reflected and will not enter the medium. Calling I_0 the intensity which actually enters, we have

$$I_0 = I - R$$

Let the medium be divided into layers of unit thickness; then the intensity I_1 entering the second layer will be some fraction q of I_0 , so that

$$I_1 = qI_0$$

The same fraction of this will be absorbed in traversing the second layer, so that

$$I_2 = qI_1 = q^2I_0$$

* Pierre Bouguer (1698–1758). Royal Professor of Hydrography at Le Havre. The law stated here is often erroneously attributed to Johann Lambert.

and

$$I_3 = q^3 I_0$$

etc. Thus we have, in general,

$$I_x = q^x I_0$$

where x is the distance through the medium and q is called the *transmission coefficient*. Figure 110 gives a plot of I_x against x for $q = 0.8$. Another equation for this curve may be derived as follows: For an infinitesimal thickness dx of the absorbing layer, the fraction dI/I of the intensity incident on it which is absorbed is proportional to dx , so that

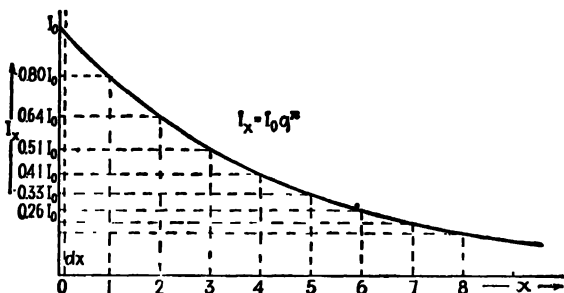


FIG. 110. Decrease in intensity of light due to absorption.

tesimal thickness dx of the absorbing layer, the fraction dI/I of the intensity incident on it which is absorbed is proportional to dx , so that

$$\frac{dI}{I} = -\alpha dx$$

To obtain the intensity I after traversing any thickness x of the medium, we may integrate both sides of this equation, determining the constant of integration by the fact that when $x = 0$, $I = I_0$, the incident intensity. This is written

$$\int_{I_0}^I \frac{dI}{I} = -\alpha \int_0^x dx$$

and gives

$$I_x = I_0 e^{-\alpha x} \quad (11v)$$

This is known as *Bouguer's exponential law* of absorption, and the constant α , which represents the fraction of the incident intensity absorbed per unit thickness when the thickness is very small, is known as the *absorption coefficient* of the medium. Since $q = e^{-\alpha}$, the absorption coefficient is equal to the negative logarithm of the transmission coefficient. These two equations expressing I_x in terms of an absorption coefficient or of the transmission coefficient are merely the same relation in different forms. In any equation representing a wave motion propa-

gated in an absorbing medium, the amplitude, which is measured by the square root of the intensity, must be multiplied by a factor $e^{-\alpha x/2}$ to take account of its decrease with x :

$$r_x = r_0 e^{-\alpha x/2}$$

11.8. Reflection and Refraction of Waves. It is a matter of common experience that when a train of waves strikes a boundary of the medium, the waves are changed in direction by reflection. In the case where the waves are incident on a boundary separating two regions where their velocity is different, they will in general be divided into a reflected train and a transmitted train whose relative intensities will depend on the magnitude of the velocity change at the boundary, on the abruptness of this change, and on the angle of incidence. If the waves approach the boundary along any direction other than the perpendicular, the trans-

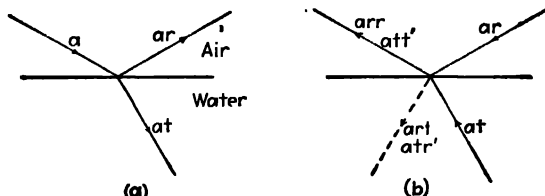


FIG. 11P. Stokes' treatment of reflection.

mitted or refracted waves will have a different direction from those incident.

These statements agree with the laws of the behavior of light rays stated in Chap. 1, since a ray represents the direction of flow of the energy of the waves, and this is usually perpendicular to the wave front (for an exception, see Sec. 25.2). The laws of reflection and refraction were deduced in Sec. 1.5 from Fermat's principle, but it is well known that they also follow from the application of Huygens' construction (Sec. 1.10) to the reflection and refraction of a plane wave.* In Fig. 11P(a), a ray incident on a plane surface of water is indicated by a , while the reflected and refracted rays are indicated by ar and at , respectively.

A question of particular interest from the standpoint of physical optics is that of a possible abrupt *change of phase* of waves when they are reflected from a boundary. For a given boundary the result will differ, as we shall now show, according to whether the waves approach from the side of higher velocity or from that of lower velocity. Thus, let the symbol a in the left-hand part of Fig. 11P represent the amplitude of a

* Sec, for example, J. K. Robertson, "Introduction to Physical Optics," 3d ed., pp. 60-67, D. Van Nostrand Company, Inc., New York,

set of waves striking the surface, let r be the fraction of the amplitude reflected, and let t be the fraction refracted (transmitted). The amplitudes of the two sets of waves will then be ar and at , as shown. Now, following a treatment given by Stokes,* imagine the two sets reversed in direction, as in part (b) of the figure. Provided there is no dissipation of energy by absorption, a wave motion is a strictly reversible phenomenon. It must conform to the law of mechanics known as the *principle of reversibility*, according to which the result of an instantaneous reversal of all the velocities in a dynamical system is to cause the system to retrace its whole previous motion. That the paths of light rays are in conformity with this principle has already been stated in Sec. 1.3. The two reversed trains, of amplitude ar and at , should accordingly have as their net effect after striking the surface a wave in air equal in amplitude to the incident wave in part (a) but traveling in the opposite direction. The wave of amplitude ar gives a reflected wave of amplitude arr and a refracted wave of amplitude art . If we call r' and t' the fractions of the amplitude reflected and refracted when the reversed wave at strikes the boundary from below, this contributes amplitudes att' and atr' to the two waves, as indicated. Now, since the resultant effect must consist only of a wave in air of amplitude a , we have

$$att' + arr = a \quad (11w)$$

and

$$art + atr' = 0 \quad (11x)$$

The second equation states that the two incident waves shall produce no net disturbance on the water side of the boundary. From Eq. 11w we obtain

$$tt' = 1 - r^2 \quad (11y)$$

and from Eq. 11x

$$r' = -r \quad (11z)$$

It might at first appear that Eq. 11y could be carried further by using the fact that intensities are proportional to squares of amplitudes and by writing, by conservation of energy, $r^2 + t^2 = 1$. This would immediately yield $t = t'$. The result is not correct, however, for two reasons. First, although the proportionality of intensity with square of amplitude holds for light traveling in a single medium, passage into a different medium

* Sir George Stokes (1819–1903). Versatile Englishman of Pembroke College, Cambridge, and pioneer in the study of the interaction of light with matter. He is known for his laws of fluorescence (Sec. 22.6) and of the rate of fall of spheres in viscous fluids. The treatment referred to here was given in his "Mathematical and Physical Papers" Vol. 2, pp. 89ff, especially p. 91.

brings in the additional factor of the index of refraction in determining the intensity. Second, it is not to the intensities that the conservation law is to be applied, but rather to the total energies of the beams. When there is a change in width of the beam, as in refraction, it must also be taken into account.

The second of Stokes' relations, Eq. 11z, shows that the reflecting power, or fraction of the intensity reflected, is the same for a wave incident from either side of the boundary, since the negative sign disappears upon squaring the amplitudes. It should be noted, however, that the waves must be incident at angles such that they correspond to angles of incidence and refraction. The difference in sign of the amplitudes in Eq. 11z indicates a difference of phase of π between the two cases, since a reversal of sign means a displacement in the opposite sense. If there is no phase change on reflection from above, there must be a phase change of π on reflection from below; or correspondingly, if there is no change on reflection from below, there must be a change of π on reflection from above.

The principle of reversibility as applied to light waves is often useful in optical problems; for example, it proves at once the interchangeability of object and image. The conclusion reached above about the change of phase is not dependent on the applicability of the principle, *i.e.*, on the absence of absorption, but holds for reflection from any boundary. It is a matter of experimental observation that in the reflection of light under the above conditions, the phase change of π occurs when the light strikes the boundary from the side of higher velocity,* so that the second of the two alternatives mentioned is the correct one in this case. A change of phase of the same type is encountered in the reflection of simple mechanical waves, such as transverse waves in a rope. Reflection with change of phase where the velocity decreases in crossing the boundary corresponds to the reflection of waves from a fixed end of a rope. Here the elastic reaction of the fixed end of the rope immediately produces a reflected train of opposite phase traveling back along the rope. The case where the velocity increases in crossing the boundary has its parallel in reflection from a free end of a rope. The end of the rope undergoes a displacement of twice the amount it would have if the rope were continuous, and it immediately starts a wave in the reverse direction having the same phase as the incident wave. We shall make use of the conclusions embodied in Eqs. 11y and 11z in discussing the interference of light (Sec. 14.1) and shall return to the question of the phase relations for reflection at any angle of incidence in Chap. 28.

* See the discussion in Sec. 13.8 under Lloyd's mirror.

*

Problems

1. Plot the equation

$$y = r \sin 2\pi \left(\frac{t}{T} - \frac{x}{\lambda} \right)$$

where $r = 6$ cm, $T = 8$ sec, $\lambda = 24$ cm, and $x = 8$ cm. Use y as ordinates and t as abscissas, extending the graph from $t = 0$ to $t = 16$ sec.

2. A point source sends out waves which, as they pass a point 100 cm from the source, generate a motion represented by the equation

$$y = r \sin \frac{2\pi}{T} t$$

where $r = 0.36$ cm and $T = \frac{1}{2}$ sec. If the waves travel at the rate of 80 cm/sec, find the equation for the motion at a point 250 cm from the source.

3. Plot the equation $y = r \sin 2\pi \left(\frac{t}{T} - \frac{x}{\lambda} \right)$, where $r = 3$ cm, $T = 6$ sec, $\lambda = 6$ cm, and $x = 10$ cm. Use y as ordinates and t as abscissas, extending the graph from $t = 0$ to $t = 12$ sec.

4. A source of plane-parallel waves vibrates according to the relation $y = r \sin (2\pi t/T)$, where $r = 2$ cm, and $T = 0.5$ sec. If the waves travel at the rate of 160 cm/sec, find the equation of motion of a particle in the wave train 760 cm from the source.

5. Find the phase difference between the vibrations of two different particles in Prob. 4 (a) at 400 cm, and (b) at 420 cm.

6. Upon measuring the photographed spectrum of a star, the lines are all found to be shifted to shorter wavelengths. A line which in a laboratory light source is at 5461.00 Å is found to be at 5460.98 Å. Assuming this shift to be a *Doppler effect*, find the velocity with which the star is approaching the earth ($c = 3 \times 10^{10}$ cm/sec).

7. If in Prob. 6 another star is observed and the same spectrum line is measured at 5461.03 Å, find the radial velocity of the star.

8. Calculate the value of the absorption coefficient α for a medium whose transmission coefficient $q = 0.6$.

9. Calculate the value of the transmission coefficient q for a medium whose absorption coefficient is 0.05 per cm.

10. Infrared light is incident on a medium whose absorption coefficient $\alpha = 0.02$ per cm. Find the decrease in intensity of the light after it has traveled 10, 20, 50, and 200 cm through the medium.

11. An underwater sound source sends out waves having a frequency of 30 kc. Find the wavelength of the waves (a) in the water, and (b) in the air above, if the velocity in water is 145,000 cm/sec and in air is 34,000 cm/sec.

12. A cord of linear density 12 g/cm extends in the x direction, and is under a tension of 5 kg of force. One end of the cord is given a transverse motion represented by $y = 2.5 \sin 12t$. Assuming that the waves keep a constant amplitude as they progress and that the other end of the cord is at infinity, find (a) the velocity of the waves, (b) the wavelength, (c) the angular phase difference between the two motions of two points along the rope 1.5 m apart.

13. A wave motion traveling along the x axis is described by the differential equation $\partial^2 y / \partial x^2 = (1/v^2) \partial^2 y / \partial t^2$. (a) Show that $y = f(x \pm vt)$ is a solution of this equation. (b) Show that $y = r \sin 2\pi[(t/T) - (x/\lambda)]$ is a particular solution.

14. Derive the relation between the transmission coefficient and the absorption coefficient.

CHAPTER 12

THE SUPERPOSITION OF WAVES

When two sets of waves are made to cross each other, as, for example, the waves created by dropping two stones simultaneously in a quiet pool, very interesting and complicated effects are observed. In the region of crossing there are places where the disturbance is practically zero, and others where it is greater than that which would be given by either wave alone. A very simple law can be used to explain these effects, which states that the resultant displacement of any point is merely the sum of the displacements due to each wave separately. This is known as the *principle of superposition* and was first clearly stated by Young* in 1802. The truth of this principle is at once evident when we observe that after the waves have passed out of the region of crossing, they appear to have been entirely uninfluenced by the other set of waves. Amplitude, frequency, and all other characteristics are just as if they had crossed an undisturbed space. This could hold only provided the principle of superposition were true. Two different observers can see different objects through the same aperture with perfect clearness, whereas the light reaching the two observers has crossed in going through the aperture. The principle is therefore applicable with great precision to light, and we may use it in investigating the disturbance in regions where two or more light waves are superimposed.

12.1. Addition of Simple Periodic Motions along the Same Line. Considering first the effect of superimposing two simple periodic waves of the same frequency, the problem resolves itself into finding the resultant motion when a particle executes two simple periodic motions at the same time. The displacements due to the two waves are here taken to be along the same line, which we shall call the y direction. If the amplitudes of the two waves are r_1 and r_2 , these will be the amplitudes of the two periodic motions impressed on the particle, and according to Eq.

* Thomas Young (1773-1829). English physician and physicist, usually called the founder of the wave theory of light. An extremely precocious child (he had read the Bible twice through at the age of four), he developed into a brilliant investigator. His work on interference constituted the most important contribution on light since Newton. His early work proved the wave nature of light but was not taken seriously by others until it was corroborated by Fresnel.

11a of the last chapter, we may write the separate displacements as follows:

$$\begin{aligned} y_1 &= r_1 \sin(\omega t + \alpha_1) \\ y_2 &= r_2 \sin(\omega t + \alpha_2) \end{aligned} \quad (12a)$$

Note that ω is the same for both waves, since we have assumed their frequencies to be the same. According to the principle of superposition, the resultant displacement y is merely the sum of y_1 and y_2 , and we have

$$y = r_1 \sin(\omega t + \alpha_1) + r_2 \sin(\omega t + \alpha_2)$$

Using the expression for the sine of the sum of two angles, this may be written

$$\begin{aligned} y &= r_1 \sin \omega t \cos \alpha_1 + r_1 \cos \omega t \sin \alpha_1 + r_2 \sin \omega t \cos \alpha_2 + r_2 \cos \omega t \sin \alpha_2 \\ &= (r_1 \sin \alpha_1 + r_2 \sin \alpha_2) \cos \omega t + (r_1 \cos \alpha_1 + r_2 \cos \alpha_2) \sin \omega t \end{aligned} \quad (12b)$$

Now, since the amplitudes r_1 and r_2 , and also the initial phases α_1 and α_2 , are constants, we are justified in setting

$$\left. \begin{aligned} r_1 \sin \alpha_1 + r_2 \sin \alpha_2 &= R \sin \theta \\ r_1 \cos \alpha_1 + r_2 \cos \alpha_2 &= R \cos \theta \end{aligned} \right\} \quad (12c)$$

provided that constant values of R and θ can be found which satisfy these equations. Squaring and adding Eqs. 12c, we have

$$\begin{aligned} R^2(\sin^2 \theta + \cos^2 \theta) &= r_1^2(\sin^2 \alpha_1 + \cos^2 \alpha_1) \\ &\quad + r_2^2(\sin^2 \alpha_2 + \cos^2 \alpha_2) + 2r_1r_2(\sin \alpha_1 \sin \alpha_2 + \cos \alpha_1 \cos \alpha_2) \end{aligned}$$

or

$$R^2 = r_1^2 + r_2^2 + 2r_1r_2 \cos(\alpha_1 - \alpha_2) \quad (12d)$$

Dividing the upper equation 12c by the lower, we obtain

$$\tan \theta = \frac{r_1 \sin \alpha_1 + r_2 \sin \alpha_2}{r_1 \cos \alpha_1 + r_2 \cos \alpha_2} \quad (12e)$$

Equations 12d and 12e show that values of R and θ exist which satisfy Eqs. 12c, and we may now rewrite Eq. 12b, substituting the right-hand members of Eqs. 12c. This gives

$$y = R \sin \theta \cos \omega t + R \cos \theta \sin \omega t$$

which has the form of the sine of the sum of two angles and can be written

$$y = R \sin(\omega t + \theta) \quad (12f)$$

This equation is the same as either of our original equations for the separate simple periodic motions but contains a new amplitude R and a new phase constant θ . Hence we have the important result that the sum of two simple periodic motions of the same frequency and along the same line is also a simple periodic motion of the same frequency. The amplitude and phase constant of the resultant motion can easily be calculated from those of the component motions by Eqs. 12d and 12e, respectively.

The addition of three or more simple periodic motions of the same frequency will likewise give rise to a resultant motion of the same type, since the motions can be added successively, each time giving an equation of the form of Eq. 12f. Unless considerable accuracy is desired, it is usually more convenient to use the graphical method described in the following section. A knowledge of the resultant phase constant θ , given by Eq. 12e, is not of much interest unless it is needed in combining the resultant motion with still another.

The resultant amplitude R depends, according to Eq. 12d, upon the amplitudes r_1 and r_2 of the component motions and upon their difference in phase $\alpha_1 - \alpha_2$. When we bring together two beams of light of equal intensity, as is done in the Michelson interferometer (Sec. 13.10), the intensity of the light at any point will be proportional to the square of the resultant amplitude. By Eq. 12d we have, in the case where $r_1 = r_2$,

$$I = R^2 = 2r^2[1 + \cos(\alpha_1 - \alpha_2)] = 4r^2 \cos^2\left(\frac{\alpha_1 - \alpha_2}{2}\right) \quad (12g)$$

If the phase difference is such that $\alpha_1 - \alpha_2 = 0, 2\pi, 4\pi, \dots$, then this gives $4r^2$, or four times the intensity of either beam. If $\alpha_1 - \alpha_2 = \pi, 3\pi, 5\pi, \dots$, the intensity is zero. For intermediate values, the intensity varies between these limits according to the square of the cosine. These modifications of intensity obtained by combining waves are referred to as *interference* effects, and we shall discuss in the next chapter several ways in which they may be brought about and used experimentally.

12.2. Vector Addition of Amplitudes. A very simple geometrical construction can be used to find the resultant amplitude and phase constant of the combined motion in the above case of two simple periodic motions along the same line. If we represent the amplitudes r_1 and r_2 by vectors making angles α_1 and α_2 with the x axis,* as in Fig. 12A(a), the resultant amplitude R is the vector sum of r_1 and r_2 , and makes an angle θ with the axis. To prove this proposition, we first note from

* Here we depart from the usual convention of measuring positive angles in the counterclockwise direction (Sec. 11.1), because it is customary in optics to represent an advance of phase by a clockwise rotation of the amplitude vector.

Fig. 12A(b) that, in the triangle formed by r_1 , r_2 , and R , the law of cosines gives

$$\begin{aligned} R^2 &= r_1^2 + r_2^2 - 2r_1r_2 \cos [\pi - (\alpha_1 - \alpha_2)] \\ &= r_1^2 + r_2^2 + 2r_1r_2 \cos (\alpha_1 - \alpha_2) \end{aligned}$$

agreeing with Eq. 12d. Furthermore the tangent of the angle θ is the ratio of the sum of the projections of r_1 and r_2 on the y axis to the sum of their projections on the x axis, so that

$$\tan \theta = \frac{r_1 \sin \alpha_1 + r_2 \sin \alpha_2}{r_1 \cos \alpha_1 + r_2 \cos \alpha_2}$$

which is the same as Eq. 12e.

If the triangle in Fig. 12A(b) is rotated clockwise about O with an angular velocity ω , the two vectors r_1 and r_2 will rotate with the same

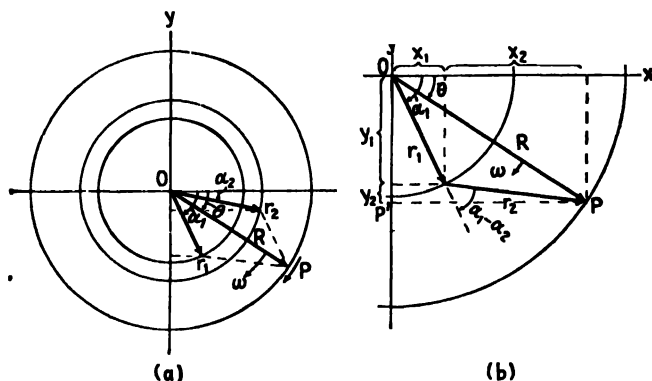


FIG. 12A. Vector addition of two amplitudes.

angular velocity, and their projections y_1 and y_2 will at every instant represent the two displacements which are to be added. The point P will then move in a circle of radius R , and its projection P' will move with the combined simple periodic motion, having always a displacement which is the sum, $y_1 + y_2$, of those due to the component motions. The position of P' corresponds to its initial position at the time $t = 0$, at which time the displacement is given by

$$OP' = y = R \sin \theta$$

As the vector R rotates, the phase angle will increase in the time t by ωt and the displacement will at any time be

$$y = R \sin (\omega t + \theta)$$

as in Eq. 12f. The phase angles of r_1 and r_2 will also increase at the same rate, so that their difference remains constant and equal to $\alpha_1 - \alpha_2$. It is convenient to have a single symbol for this phase difference, which we shall denote δ . It is this difference which is important, rather than the actual phase of either motion, since by Eq. 12d, δ determines the resultant intensity by the relation

$$R^2 = r_1^2 + r_2^2 + 2r_1r_2 \cos \delta \quad (12h)$$

The graphical method is particularly useful where we have more than two motions to compound. Figure 12B shows the result of adding five motions of equal amplitudes r and having equal phase differences δ . Clearly the intensity $I = R^2$ can here vary between zero and $25r^2$, according to the phase difference δ . This is the problem which arises in finding the intensity pattern from a diffraction grating, as discussed in Chap. 17. The five equal amplitudes shown in the figure might be contributed by five apertures of a grating, an instrument which has as its primary purpose the introduction of an equal phase difference in the light from each successive pair of apertures. It will be noted that as Fig. 12B is drawn the vibrations, starting with that at the origin, lag successively farther *behind* in phase.

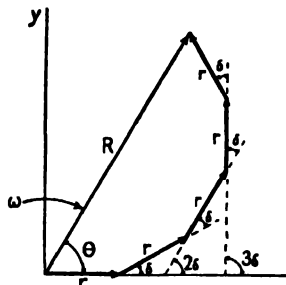


FIG. 12B. Vector addition of five equal amplitudes having the same magnitude and phase differences δ .

12.3. Superposition of Two Wave Trains of the Same Frequency. From the preceding section we may conclude directly that the result of superimposing two trains of simple periodic waves of the same frequency traveling along the same line will be another simple periodic wave of that frequency, but having a new amplitude which is determined for given values of r_1 and r_2 by the phase difference δ between the two motions imparted to any particle by the two waves. As an example, let us find the resultant wave produced by two waves of equal frequency and amplitude traveling in the same direction $+x$, but with one a distance Δ ahead of the other. The equations of the two waves, in the form of Eq. 11j, will be

$$y_1 = r \sin 2\pi \left(\frac{t}{T} - \frac{x}{\lambda} \right) \quad (12i)$$

$$y_2 = r \sin 2\pi \left(\frac{t}{T} - \frac{x + \Delta}{\lambda} \right) \quad (12j)$$

Adding these two equations, we find the resultant wave to be given by

$$y = y_1 + y_2 = r \left[\sin 2\pi \left(\frac{t}{T} - \frac{x}{\lambda} \right) + \sin 2\pi \left(\frac{t}{T} - \frac{x + \Delta}{\lambda} \right) \right]$$

From the trigonometric equation for the sum of the sines of two angles,

$$\sin A + \sin B = 2 \sin \frac{1}{2}(A + B) \cos \frac{1}{2}(A - B) \quad (12k)$$

we get

$$y = r \left[2 \sin \pi \left(\frac{t}{T} - \frac{x}{\lambda} + \frac{t}{T} - \frac{x + \Delta}{\lambda} \right) \cos \pi \left(\frac{t}{T} - \frac{x}{\lambda} - \frac{t}{T} + \frac{x + \Delta}{\lambda} \right) \right]$$

or

$$y = 2r \cos \frac{\pi \Delta}{\lambda} \sin 2\pi \left(\frac{t}{T} - \frac{x + \frac{\Delta}{2}}{\lambda} \right) \quad (12l)$$

This corresponds to a new wave of the same frequency but with the amplitude $2r \cos (\pi \Delta / \lambda)$. When Δ is a small fraction of a wavelength, this amplitude will be nearly $2r$, while if Δ is in the neighborhood of

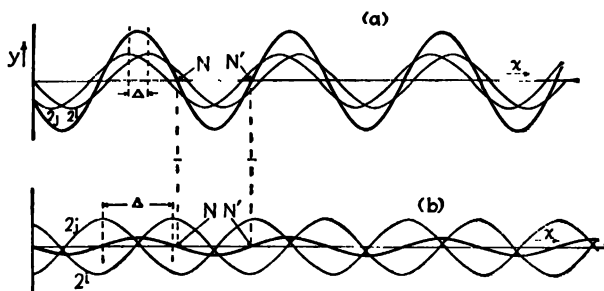


FIG. 12C. (a) Superposition of two wave trains almost in phase. (b) Superposition of two wave trains almost 180° out of phase.

$\frac{1}{2} \lambda$, it will be practically zero. These cases are illustrated in Fig. 12C, where the waves represented by Eqs. 12i and 12j (light curves) and 12l (heavy curve) are plotted at the time $t = 0$. In these figures it will be noted that the algebraic sum of the ordinates of the light curves at any value of x equals the ordinate of the heavy curve. The student may easily verify by such graphical construction the facts that the two amplitudes need not necessarily be equal to obtain a sine wave as the resultant and that the addition of any number of waves of the same frequency and wavelength also gives a similar result. In any case, the resultant wave form will have a constant amplitude, since the component waves and their resultant all move with the same velocity and maintain the same relative position. The true state of affairs may be pictured by having all the waves in Fig. 12C move toward the right with a given velocity.

The formation of the so-called "standing waves" in a vibrating cord, giving rise to nodes and loops, is an example of the superposition of two wave trains of the same frequency and amplitude but traveling in opposite directions. A wave in a cord is reflected from the end, and the direct and reflected waves must be added to obtain the resultant motion of the cord. Two such waves may be represented by the equations

$$y_1 = r \sin 2\pi \left(\frac{t}{T} - \frac{x}{\lambda} \right)$$

$$y_2 = r \sin 2\pi \left(\frac{t}{T} + \frac{x}{\lambda} \right)$$

By addition one obtains, in the same manner as for Eq. 12I,

$$y = 2r \cos \left(-\frac{2\pi x}{\lambda} \right) \sin 2\pi \frac{t}{T}$$

which represents the standing waves. For any value of x we have a simple periodic motion, whose amplitude varies with x between the limits $2r$ for $x = \lambda/2, \lambda, 3\lambda/2, 2\lambda, \dots$ and zero for $x = \lambda/4, 3\lambda/4, 5\lambda/4, \dots$. The first set of points correspond to the antinodes, where the motion is a maximum, and the second set to the nodes, where there is no motion. Both the nodes and the antinodes are spaced one-half wavelength apart. Figure 12C(a) and (b) may also serve to illustrate this case if we imagine the two light curves to be moving in opposite directions. The resultant curve, instead of moving unchanged toward the right, now oscillates between a straight-line position at the times $t = 0, T/2, T, 2T/2, \dots$ and a sine curve with amplitude $2r$ at the times $t = T/4, 3T/4, \dots$. Nodes occur, for example, at N and N' .

The standing waves produced by reflecting light at normal incidence from a polished mirror may be observed by means of an experiment due to Wiener,* which is illustrated in Fig. 12D. A specially prepared photographic film only one-twentieth of a wavelength thick is placed in an inclined position in front of the reflecting surface so that it will cross the loops and nodes successively, as at $A, a, B, b, C, c, D, d, \dots$. The light will affect the plate only where there is an appreciable amount of vibration, and not at all at the nodes. As expected, the developed plate

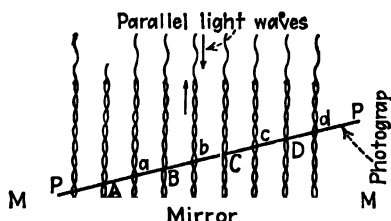


FIG. 12D. Wiener's experiment of standing light waves produced by ordinary reflection.

* O. Wiener, *Ann. Physik*, 40, 203, 1890.

showed a system of dark bands, separated by lines of no blackening where it crossed the nodes. Decreasing the angle of inclination of the plate with the reflecting surface caused the bands to move farther apart, since a smaller number of nodal planes are cut in a given distance. On measuring these bands, an important fact was established: the standing waves have a node at the reflecting surface. The phase relations of the direct and reflected waves at this point are therefore such that they continuously annul each other. This is analogous to the reflection of the waves in a rope from a fixed end. Other experiments of a similar nature were performed by Wiener and these will be discussed more in detail in Sec. 28.11.

12.4. Superposition of Many Waves with Random Phases. Suppose that we now consider a large number of wave trains of the same frequency and amplitude to be traveling in the same direction, and specify that the amount by which each train is ahead or behind any other is a matter of pure chance. From what has been said

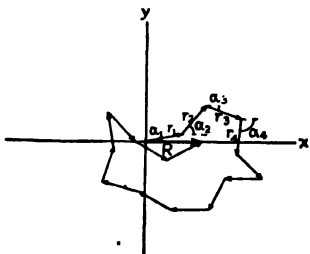


FIG. 12E. Superposition of many waves with random phases.

above, we can conclude that the resultant wave will be another simple periodic wave of the same frequency, and it becomes of interest to inquire as to the amplitude and intensity of this wave. Let the individual amplitudes be r , and let there be n wave trains superimposed. The amplitude of the resultant wave will be the amplitude of motion of a particle undergoing n simple periodic motions at once, each of amplitude r . If these motions were all in the same phase, the resultant amplitude would be nr , and the intensity n^2r^2 , or n^2 times that of one wave. In the case we are considering, however, the phases are distributed purely at random. If one were to use the graphical method of compounding amplitudes (Sec. 12.2), he would now obtain a picture like Fig. 12E. The phases $\alpha_1, \alpha_2, \dots$ take perfectly arbitrary values between 0 and 2π . The intensity due to the superposition of such waves will now be determined by the square of the resultant R . To find R^2 we must square the sum of the projections of all vectors r on the x axis and add the square of the corresponding sum for the y axis. The sum of the x projections is

$$r(\cos \alpha_1 + \cos \alpha_2 + \cos \alpha_3 + \dots + \cos \alpha_n)$$

When the quantity in parentheses is squared, we obtain terms of the form $\cos^2 \alpha_1$ and others of the form $2 \cos \alpha_1 \cos \alpha_2$. When n is large, one

might expect the latter terms to cancel out, because they take both positive and negative values. In any *one* arrangement of the vectors this is far from true, however, and in fact the sum of these cross-product terms actually increases approximately in proportion to their number. Thus we do not obtain a definite result with one given array of randomly distributed waves. In computing the intensity in any physical problem, we are always presented with a large number of such arrays, and we wish to find their average effect. In this case it is safe to conclude that the cross-product terms will average to zero, and we have only the $\cos^2 \alpha$ terms to consider. Similarly, for the y projections of the vectors one obtains $\sin^2 \alpha$ terms, and the terms like $2 \sin \alpha_1 \sin \alpha_2$ cancel. Therefore we have

$$I = R^2 = r^2(\cos^2 \alpha_1 + \cos^2 \alpha_2 + \cos^2 \alpha_3 + \cdots + \cos^2 \alpha_n) \\ + r^2(\sin^2 \alpha_1 + \sin^2 \alpha_2 + \sin^2 \alpha_3 + \cdots + \sin^2 \alpha_n)$$

Now since $\sin^2 \alpha_k + \cos^2 \alpha_k = 1$, we find at once that

$$I = r^2 \times n$$

Thus the average intensity resulting from the superposition of n waves with random phases is just n times that due to a single wave. This means that the amplitude R in Fig. 12E, instead of averaging to zero when a large number of vectors r are repeatedly added in random directions, must actually increase in length as n increases, being proportional to \sqrt{n} .

The above considerations may be used to explain why, when a large number of violins in an orchestra are playing the same note, interference between the sound waves need not be considered. Owing to the random condition of phases, 100 violins would give about 100 times the intensity due to one alone. The atoms in a sodium flame are emitting light without any systematic relation of phases, and furthermore each is shifting its phase many million times per second. Thus we may safely conclude that the observed intensity is that due to one atom multiplied by the number of atoms.

12.5. Complex Waves. The waves we have considered so far have been of the simple periodic type, in which the displacements at any instant are represented by a sine curve. As we have seen, superposition of any number of such waves having the same frequency, but arbitrary amplitudes and phases, still gives rise to a resultant wave of this simple type. However, if only two waves having appreciably different frequencies are superimposed, the resulting wave is complex; *i.e.*, the motion of one particle is no longer simple periodic motion, and the wave contour

is not a sine curve. The analytical treatment of such waves will be referred to in the following section, and here we shall consider only some of their more qualitative aspects.

It is instructive to examine the results of adding graphically two or more waves traveling along the same line and having various relative frequencies, amplitudes, and phases. The wavelengths are determined by the frequencies according to the relation $\nu\lambda = v$, so that greater frequency means shorter wavelength, and vice versa. Figure 12*F* illustrates the addition for a number of cases, the resultant curves in each case being obtained, according to the principle of superposition, by merely adding

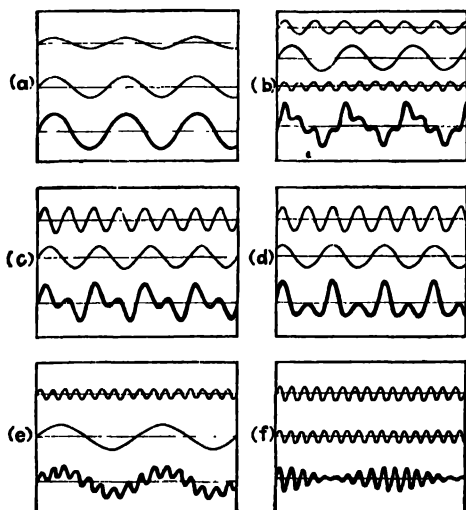


FIG. 12*F*. Superposition of two or more waves traveling in the same direction with different relative frequencies, amplitudes, and phases.

algebraically the displacements due to the individual waves at every point. Figure 12*F*(a) illustrates the case, mentioned in Sec. 12.3, of the addition of two waves of the same frequency but different amplitudes. The resultant amplitude depends on the phase difference, which in the figure is taken as zero. Other phase differences would be represented by shifting one of the component waves laterally with respect to the other and will give a smaller amplitude for the resultant sine wave, its smallest value being the difference in the amplitudes of the components. In (b) three waves of different frequencies, amplitudes, and phases are added, giving a complex wave as the resultant, which is evidently very different from a simple periodic curve. In (c) and (d), where two waves of the same amplitude but frequencies in the ratio 2:1 are added, it is seen that changing the phase difference may produce a resultant of very different

form. If these represent sound waves, the eardrum would actually vibrate in a manner represented by the resultant in each case, yet the ear mechanism would respond to two frequencies and these would be heard and interpreted as the two original frequencies regardless of their phase difference. If the resultant wave *fo.* is represent visible light, the eye would receive the same sensation of a mixture of two colors, regardless of the phase difference. Finally (*e*) shows the effect of adding a wave of very high frequency to one of very low frequency, and (*f*) the effect of adding two of nearly the same frequency. In the latter case, the resultant wave divides up into groups, which in sound produce the well-known phenomenon of beats. In any of the above cases, if the component waves all travel with the same velocity, the resultant wave form will evidently move with this velocity, keeping its contour unchanged.

Experimental illustrations of the superposition of waves are easily accomplished with the apparatus shown in Fig. 12*G*. Two small mirrors, M_1 and M_2 , are cemented to thin strips of spring steel which are clamped vertically and illuminated by a narrow beam of light. Such a beam is conveniently produced by a single-filament automobile-headlight bulb S and a lens L to focus an image of it on the screen. This beam is reflected in succession from the two mirrors, and if one of them is set vibrating, the reflected beam will vibrate up and down with simple periodic motion. If now this beam on its way to the screen is reflected from a rotating mirror, the spot of light will trace out a sine wave form which will appear continuous by virtue of the persistence of vision. When both M_1 and M_2 are set vibrating at once, the resultant wave form is the superposition of that produced by each separately. In this way all the curves of Fig. 12*F* may be produced by using two or more strips of suitable frequencies. The frequencies may be easily altered by changing the free length of the strips above the clamps.

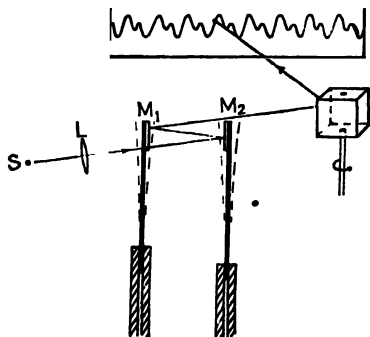


FIG. 12*G*. Mechanical and optical arrangement for illustrating the superposition of two waves.

Since for visible light the frequency determines the color, complex waves of light are produced when beams of light of different colors are used. The "impure" colors which are not found in the spectrum will

therefore have waves of a complex form. White light, which, since Newton's original experiments with prisms, we usually speak of as composed of a mixture of all colors, is the extreme example of the superposition of a great number of waves having frequencies differing only by infinitesimal amounts. We shall discuss the resultant wave form for white light in the following section. Strictly speaking, as mentioned in Sec. 11.3, it is impossible to produce light consisting of simple periodic waves, so that even the most nearly monochromatic light that we can produce must still be regarded as containing a finite range of frequencies.

12.6. Fourier Analysis. Since we may build up a wave of very complex form by the superposition of a number of simple waves, it is of interest to ask to what extent the converse processes may be accomplished—that of decomposing a complex wave into a number of simple ones. According to a theorem due to Fourier, any *periodic* function may be represented as the sum of a number of sine and cosine functions. By a periodic function we mean one which repeats itself exactly in successive equal intervals, such as the lower curve in Fig. 12F(b). The wave is given by an equation of the type

$$y = r_0 + r_1 \sin(\omega t + \alpha_1) + r_2 \sin(2\omega t + \alpha_2) + r_3 \sin(3\omega t + \alpha_3) + \dots \quad (12m)$$

This is known as a *Fourier series* and contains, beside the constant term r_0 , a series of terms having amplitudes r_1, r_2, \dots and frequencies $\omega/2\pi, 2\omega/2\pi, 3\omega/2\pi, \dots$. This means that the resultant wave is regarded as built up of a number of waves whose wavelengths are as $1:\frac{1}{2}:\frac{1}{3}:\frac{1}{4}, \dots$. In the case of sound, these represent the fundamental note and its various harmonics. The evaluation of the amplitude coefficients r_1, r_2, \dots for a given wave form can be carried out by a straightforward mathematical process in the case of some fairly simple wave forms, but in general this is a difficult matter. Usually one must have recourse to one of the various forms of "harmonic analyzer," a mechanical device for determining the amplitudes and phases of the fundamental and its harmonics.*

Fourier analysis is not often of direct use in studying light waves, because it is impossible to observe directly the form of a light wave. For sound this can be done, and it is in the investigation of the quality of sounds that Fourier analysis has been most extensively used. However, it is important for us to understand the principles of the method,

*For a detailed account of harmonic analyzers see D. C. Miller, "The Science of Musical Sounds," The Macmillan Company, New York, 1922. A good discussion of Fourier analysis may be found in E. H. Barton, "Textbook of Sound," 1st ed., p. 83ff, Macmillan & Co., Ltd., London, 1908.

because, as we shall see, a grating or a prism essentially performs a Fourier analysis of the incident light, revealing the various component frequencies which it contains and which appear as *spectral lines*.

Fourier analysis is not limited to waves of a periodic character. The upper part of Fig. 12H shows three types of waves which are not periodic, because, instead of repeating their contour indefinitely, the waves have zero displacement beyond a certain finite range. Hence such groups cannot be represented by Fourier series, but instead *Fourier integrals* must be used, in which the component waves differ only by infinitesimal increments of wavelength. By suitably distributing the amplitudes for the various components, any arbitrary wave form may be expressed by such an integral. The three lower curves in Fig. 12H represent qualitatively the frequency distribution of the amplitudes which will produce the corresponding wave groups shown above. That is, the upper curves

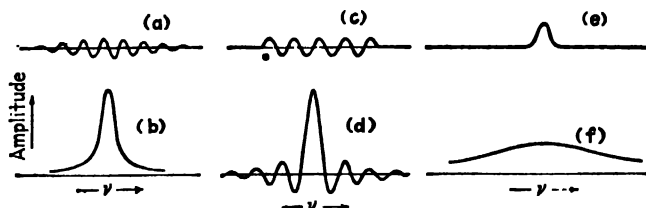


FIG. 12H. Distribution of amplitudes in different frequencies for various types of wave disturbance of finite length.

represent the actual wave contour of the group, and this contour may be synthesized by adding up a very large (strictly, an infinite) number of wave trains, each of frequency differing only infinitesimally from the next. The curves shown immediately below each group show the necessary amplitudes of the components of each frequency, in order that their superposition may produce the wave form indicated above.

Curve (a) shows a typical *group* of waves, such as is involved in the light which forms a spectral line of finite width. The group of curve (c) would be produced by passing perfectly monochromatic light through a shutter which is open for an extremely short time. It is worth remarking here that the corresponding amplitude distribution, shown in curve (d), is exactly that obtained for the Fraunhofer diffraction from a single slit and described in Sec. 15.3. Another interesting case, shown in curve (e), is that of a single *pulse*, such as the sound pulse sent out by a pistol shot or (better) the discharge of a spark. The form of such a pulse may resemble that shown, and when a Fourier analysis of it is made, it yields a continuous distribution of wavelengths as shown in curve (f). For light, such a distribution is called a *continuous spectrum*, and is obtained with sources of white light such as an incandescent solid. The distri-

bution of intensity in different wavelengths, which is proportional to the square of the ordinates in the curve, is determined by the exact shape of the pulse. This view of the nature of white light is one which has been emphasized by Gouy and others,* and raises the question as to whether Newton's experiments on refraction by prisms, which are usually said to prove the composite nature of white light, were of much significance in this respect. Since white light may be regarded as consisting merely of a succession of random pulses, of which the prism performs a Fourier analysis, the view that the colors are manufactured by the prism, which was held by Newton's predecessors, may be regarded as equally correct.

12.7. Group Velocity. It will be readily seen that, if all the component simple waves making up a group travel with the same velocity, the group will move with this velocity and maintain its form unchanged. If, however, the velocities vary with wavelength, this is no longer true, and the group will change its form as it progresses. This situation exists for water waves, and if one watches the individual waves in the group sent out by dropping a stone in still water, they will be found to be moving faster than the group as a whole, dying out at the front of the group and reappearing at the back. Hence in this case the group velocity is less than the wave velocity, a relation which always holds when the velocity of longer waves is greater than that of shorter ones. It is important to establish a relation between the group velocity and wave velocity, and this can easily be done by considering the groups formed by superimposing two waves of slightly different wavelength, such as those already discussed and illustrated in Fig. 12*F(f)*. We shall suppose that the two waves have equal amplitudes r , that their wavelengths λ and $\lambda + d\lambda$ differ only slightly, and that the velocities are v and $v + dv$, respectively. The resultant wave may then be written as the sum of the two, using the wave equation 11*f*, which is the more convenient in this case. We have

$$y = r \sin \frac{2\pi}{\lambda} (vt - x) + r \sin \frac{2\pi}{\lambda + d\lambda} [(v + dv)t - x]$$

which may, by use of Eq. 12*k*, be written

$$y = 2r \sin \pi \left[t \left(\frac{v}{\lambda} + \frac{v + dv}{\lambda + d\lambda} \right) - x \left(\frac{1}{\lambda} + \frac{1}{\lambda + d\lambda} \right) \right] \\ \cos \pi \left[t \left(\frac{v}{\lambda} - \frac{v + dv}{\lambda + d\lambda} \right) - x \left(\frac{1}{\lambda} - \frac{1}{\lambda + d\lambda} \right) \right]$$

* The reader will find the more detailed discussion of the various representations of white light given in R. W. Wood, "Physical Optics" 1st or 2d ed., The Macmillan Company, New York, of interest in this connection.

Since when $d\lambda$ is small compared to λ we can neglect the difference between $\lambda(\lambda + d\lambda)$ and λ^2 , we can write

$$\frac{1}{\lambda} - \frac{1}{\lambda + d\lambda} = \frac{d\lambda}{\lambda^2} \quad \text{and} \quad \frac{v}{\lambda} - \frac{v + dv}{\lambda + d\lambda} = \frac{v d\lambda - \lambda dv}{\lambda^2}$$

so that the complete expression becomes

$$y = 2r \sin \frac{2\pi}{\lambda} (vt - x) \cos \frac{\pi d\lambda}{\lambda^2} \left(v \frac{d\lambda - \lambda dv}{d\lambda} t - x \right) \quad (12n)$$

Figure 12I(c) is a plot of this equation at the time $t = 0$. The solid line has a wavelength λ corresponding to the sine factor, while the dotted line corresponds to the cosine factor, with a wavelength of $2\lambda^2/d\lambda$. When the two factors are multiplied together, the amplitude varies in the way shown, the waves dividing up into groups. The velocity of any one group is the velocity with which the maximum of the group moves and will be different from (here, less than) the velocity of the individual waves. The

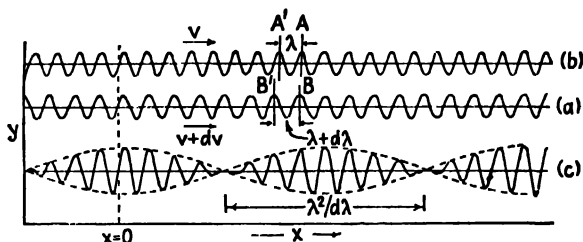


FIG. 12I. Illustrating groups and group velocity of two waves of slightly different wavelength and velocity.

group velocity is the velocity of the cosine factor, which is seen from Eq. 12n to be

$$u = \frac{v d\lambda - \lambda dv}{d\lambda}$$

Hence the relation between the *group velocity* u and the *wave velocity* v is

$$u = v - \lambda \frac{dv}{d\lambda} \quad (12o)$$

This equation, although derived for an especially simple type of group, is quite general and can be shown to hold for any group whatever, as, for example, the three illustrated in Fig. 12H(a), (c), and (e). Equation 12o can also be derived in a less mathematical way by considering the motions of the two component wave trains shown in Fig. 12I(a) and (b). At the instant shown, the crests A and B of the two trains coincide to produce

a maximum for the group. A little later the faster waves will have gained a distance $d\lambda$ on the slower ones, so that B' coincides with A' , and the maximum of the group will have moved back a distance λ . Since the difference in velocity of the two trains is dv , the time required for this is $d\lambda/dv$. But in this time both wave trains have been moving to the right, the upper one moving a distance $v d\lambda/dv$. The net displacement of the maximum of the group is thus $v (d\lambda/dv) - \lambda$ in the time $d\lambda/dv$, so that we obtain, for the group velocity,

$$u = \frac{v (d\lambda/dv) - \lambda}{d\lambda/dv} = v - \lambda \frac{dv}{d\lambda}$$

in agreement with Eq. 12o.

A picture of the groups formed by two waves of slightly different frequency may easily be produced with the apparatus described in Sec. 12.5. It is merely necessary to adjust the two vibrating strips until the frequencies differ by only a few vibrations per second.

The group velocity is important for light, since it is the only velocity which we can observe experimentally. We know of no means of following the progress of an individual wave in a group of light waves; instead, we are obliged to measure the rate at which a wave train of finite length conveys the *energy*, a quantity which can be observed. The wave and group velocities become the same in a medium which has no dispersion, i.e., in which $dv/d\lambda = 0$, so that waves of all lengths travel with the same speed. As was pointed out in Sec. 11.4, this is accurately true for light traveling in a vacuum, so that there is no difference between group and wave velocities in this case.

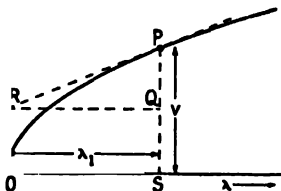


FIG. 12J. Graphical determination of group velocity from a wave velocity curve.

12.8. Graphical Relation between Wave and Group Velocity. There is a very simple geometrical construction by which we may determine the group velocity from a curve of the wave velocity against wavelength.

It is based upon the graphical interpretation of Eq. 12o. As an example, the curve of Fig. 12J represents the variation of the wave velocity with λ for water waves in deep water (gravity waves) and is drawn according to the theoretical equation $v = \text{const.} \times \sqrt{\lambda}$. At a certain wavelength λ_1 , the waves have a velocity v , and the slope of the curve at the corresponding point P gives $dv/d\lambda$. The line PR , drawn tangent to the curve

at this point, intersects the v axis at a point R , the ordinate of which is the group velocity u for waves of wavelength in the neighborhood of λ_1 . This is evident from the fact that PQ equals $\lambda_1 dv/d\lambda$, i.e., the abscissa of P multiplied by the slope of PR . Hence QS , which is drawn equal to RO , represents the difference $v - \lambda_1 dv/d\lambda$, and this is just the value of u , by Eq. 12o. In the particular example chosen here, it will be left as a problem for the student to prove that $u = \frac{1}{2}v$ for any value of λ . In water waves of this type, the individual waves therefore move with twice the velocity with which the group as a whole progresses.

12.9. Addition of Simple Periodic Motions at Right Angles. Consider the effect when two simple periodic motions of the same frequency but having displacements in two perpendicular directions are impressed simultaneously on a point. Choosing the directions as y and z , we may express the two motions as follows:

$$\left. \begin{aligned} y &= r_1 \sin(\omega t + \alpha_1) \\ z &= r_2 \sin(\omega t + \alpha_2) \end{aligned} \right\} \quad (12p)$$

These are to be added, according to the principle of superposition, to find the path of the resultant motion. This may be found by eliminating t from the two equations

$$\frac{y}{r_1} = \sin \omega t \cos \alpha_1 + \cos \omega t \sin \alpha_1 \quad (12q)$$

$$\frac{z}{r_2} = \sin \omega t \cos \alpha_2 + \cos \omega t \sin \alpha_2 \quad (12r)$$

Multiplying Eq. 12q by $\sin \alpha_2$ and Eq. 12r by $\sin \alpha_1$ and subtracting the first equation from the second, there results

$$-\frac{y}{r_1} \sin \alpha_2 + \frac{z}{r_2} \sin \alpha_1 = \sin \omega t (\cos \alpha_2 \sin \alpha_1 - \cos \alpha_1 \sin \alpha_2) \quad (12s)$$

Similarly, multiplying Eq. 12q by $\cos \alpha_2$ and Eq. 12r by $\cos \alpha_1$, and subtracting the second from the first, we obtain

$$\frac{y}{r_1} \cos \alpha_2 - \frac{z}{r_2} \cos \alpha_1 = \cos \omega t (\cos \alpha_2 \sin \alpha_1 - \cos \alpha_1 \sin \alpha_2) \quad (12t)$$

We may now eliminate t from Eqs. 12s and 12t by squaring and adding these equations. This gives

$$\sin^2(\alpha_1 - \alpha_2) = \frac{y^2}{r_1^2} + \frac{z^2}{r_2^2} - \frac{2yz}{r_1 r_2} \cos(\alpha_1 - \alpha_2) \quad (12u)$$

as the equation for the resultant path. In Fig. 12K the heavy curves are graphs of this equation for various values of the phase difference $\delta = \alpha_1 - \alpha_2$. Except for the special cases where they degenerate into straight lines, these curves are all ellipses. The principal axes of the ellipse are in general inclined to the y and z axes but coincide with them when $\delta = \pi/2, 3\pi/2, 5\pi/2, \dots$, as can readily be seen from Eq. 12u. In this case

$$\frac{y^2}{r_1^2} + \frac{z^2}{r_2^2} = 1$$

which is the equation of an ellipse with semiaxes r_1 and r_2 , coinciding with

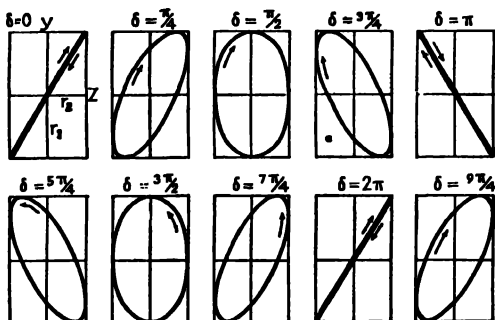


FIG. 12K. Composition at right angles of two simple periodic motions of the same frequency but different phase.

the y and z axes, respectively. When $\delta = 0, 2\pi, 4\pi, \dots$, we have

$$y = \frac{r_1}{r_2} z$$

representing a straight line passing through the origin, with a slope r_1/r_2 . If $\delta = \pi, 3\pi, 5\pi, \dots$,

$$y = -\frac{r_1}{r_2} z$$

a straight line with the same slope, but of opposite sign.

That the two cases $\delta = \pi/2$ and $\delta = 3\pi/2$, although giving the same path, are physically different types of motion may be seen from the graphical construction of Fig. 12L. This illustrates a convenient way of finding the resultant of the two motions at right angles, using the projection of uniform circular motion discussed in Sec. 11.1 for generating each of the two simple periodic motions. The actual motions to be combined are along the vertical and horizontal lines labeled y and z , respectively, and are projected into a rectangle of sides $2r_1$ and $2r_2$ to obtain the resultant motion. Dividing a complete period into eighths,

the numbers 0 to 8 represent the positions of the generating points, and the resultant position, at these intervals. The phase difference is determined by the starting points, and in the figure it will be seen that for $\delta = \pi/2$ the z motion starts a quarter period behind, since $z = -r_2$ when $y = 0$ (and *increasing*). Similarly, with $\delta = 3\pi/2$, $z = +r_2$ for $y = 0$. The difference between the two cases is seen to be in the direction of rotation in the ellipse. If this procedure is carried out for the other values of δ , it will be found that the motions are in the directions shown by the arrows in Fig. 12K. The student will find it instructive to make this construction for other cases, including those in which the frequency of one motion is twice or three times the other (or in general with the

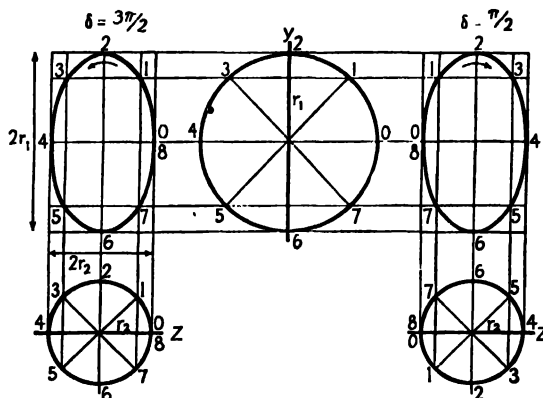


FIG. 12L. Graphical composition of two simple periodic motions at right angles.

frequencies in the ratio of any two whole numbers), which gives rise to interesting curves known as *Lissajous figures*.

Of these Lissajous figures, only those illustrated in Fig. 12K, for equal frequencies, are of particular interest for the study of light. The so-called *plane-polarized* light (Chap. 24) approximates a sine wave lying in a plane—say the x,y plane of Fig. 12M—and the displacements are linear displacements in the y direction. If one combines a beam of this light with another consisting of plane-polarized waves lying in the x,z plane (dotted curve) and having a constant phase difference with the first, the resultant motion at any value of x will be a certain ellipse in the y,z plane. Such light is said to be *elliptically polarized* and may readily be produced by various means (Chap. 26). A special case occurs when the amplitudes r_1 and r_2 of the two waves are equal and the phase difference is an odd multiple of $\pi/2$. The vibration form is then a circle, and the light is said to be *circularly polarized*. When the direction of rotation

is clockwise ($\delta = \pi/2, 5\pi/2, \dots$) looking opposite to the direction in which the light is traveling, the light is called *right* circularly polarized, while if the rotation is counterclockwise ($\delta = 3\pi/2, 7\pi/2, \dots$), it is called *left* circularly polarized.

Lissajous figures of all kinds are readily demonstrated with the apparatus described in Sec. 12.5. For this, the two strips are arranged to vibrate at right angles to each other, and the rotating mirror is dispensed with. Thus one strip imparts a horizontal vibration to the spot of light, and the other a vertical vibration, while both at once give a certain Lissajous figure. The figure will remain fixed if the frequencies are exactly equal, or in the ratio of simple whole numbers. If the frequencies

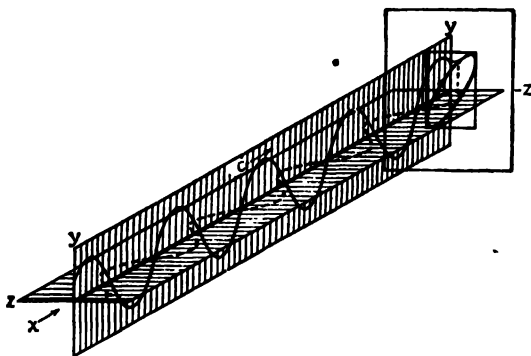


FIG. 12M. Composition of two simple periodic waves at right angles.

are only slightly different from these, the figure will progress through all its forms characteristic of various values of the phase difference, such as those shown in Fig. 12K.

Problems

1. Calculate the resultant amplitude R and the initial phase angle θ for the sum of the following three motions: $y_1 = 3 \sin (\pi/8)t$, $y_2 = 3 \sin [(\pi/8)t + (\pi/4)]$, $y_3 = 3 \sin [(\pi/8)t + (\pi/2)]$.
2. Repeat Prob. 1 for motions: $y_1 = 3 \sin [(\pi/4)t + (\pi/2)]$, $y_2 = 4 \sin [(\pi/4)t + \pi]$, $y_3 = 3 \sin [(\pi/4)t + (3\pi/2)]$.
3. Plot the two equations $y_1 = 2 \sin (\pi/3)t$ and $y_2 = 3 \sin [(\pi/3)t + (5\pi/6)]$, and their resultant $y = R \sin [(\pi/3)t + \theta]$. Use t as abscissas and determine the ordinates y by adding.
4. Calculate the resultant R and the initial phase angle θ in Prob. 3.
5. Two waves vibrating in the same plane and traveling through the same medium

are given by the equations $y_1 = r_1 \sin (2\pi/\lambda_1)(v_1 t - x)$ and $y_2 = r_2 \sin (2\pi/\lambda_2)(v_2 t - x)$, where $r_1 = 1$ cm, $r_2 = 1$ cm, $\lambda_1 = 4.0$ cm, $\lambda_2 = 4.1$ cm, $v_1 = 100$ cm/sec, and $v_2 = 101$ cm/sec. Calculate the group velocity.

6. Two periodic motions at right angles are given by the equations $y = 3 \sin [(\pi/4)t + (\pi/4)]$, $z = 5 \sin [(\pi/4)t + (3\pi/4)]$. Plot the resultant motion as in Fig. 12L. Indicate the starting point and the direction of motion.

7. Draw the Lissajous figure resulting from the composition of two simple periodic motions at right angles whose analytical expressions are

$$x = 4 \sin \left(\frac{\pi}{2} t + \frac{\pi}{4} \right) \quad \text{and} \quad y = 6 \sin \left(\pi t + \frac{\pi}{2} \right)$$

8. Two periodic motions in the same line are given by $x_1 = 3 \sin \frac{\pi}{2} t$ and $x_2 = 2 \cos \left(\frac{\pi}{2} t + \frac{\pi}{4} \right)$. Find the amplitude and initial phase of the resultant motion. What is the equation of the resultant motion?

9. Standing waves are produced, as in Wiener's experiment, by reflecting light normally from a plane mirror. If the light has a wavelength of 6563 Å, find the number of dark bands per centimeter on the photographic plate when the film is inclined (a) at 1° to the reflecting surface, and (b) at 10° .

10. The velocity of very short water waves is given by $v = (2\pi T/\lambda d)^{1/2}$ where T is the surface tension and d the density. Calculate (a) the wave velocity for waves 0.5 cm long, and (b) the group velocity for a group of these waves.

11. Two waves traveling in opposite directions in the same medium are symbolized by the equations $y_1 = 5 \sin \pi(5t - x)$ and $y_2 = 5 \sin \pi(5t + x)$. Plot the resultant wave form between $x = 2$ and $x = 6$ at the times $t = \frac{1}{4}$ sec and $t = \frac{3}{4}$ sec.

12. Two trains of waves traveling in the same direction have amplitudes of 4 cm and 3 cm respectively. They have equal wavelengths of 10 cm and equal velocities of 160 cm/sec. The crests of the 3-cm wave are a third of a wavelength behind those of the 4-cm wave. (a) Plot the resultant wave form by graphical addition. (b) Write an equation for the resultant wave.

13. Plot y against t for the Fourier series

$$y = 10 + 8 \sin \pi t + 4 \sin 2\pi t + 2 \sin 3\pi t$$

14. Derive an equation for the elliptical vibration in Prob. 6.

15. As suggested in Sec. 12.8, prove that for gravity water waves, for which $v = k\sqrt{\lambda}$, the group velocity $u = \frac{1}{2}v$.

16. The velocity of surface waves (gravity waves) in deep water has been shown to be given by $v = \sqrt{g\lambda/2\pi}$. Plot v against λ for waves from $\lambda = 0$ to 100 cm, and graphically determine the group velocity for $\lambda = 25$ cm, 50 cm, and 75 cm.

17. Differentiate the equation for the velocity of water waves in Prob. 16 and calculate the group velocity for $\lambda = 25$ cm, 50 cm, and 75 cm.

18. The velocity of surface waves on liquids, controlled by gravity and surface tension, is given by

$$v = \sqrt{\frac{\lambda}{2\pi} \left(g + \frac{4\pi^2 \gamma}{\lambda^3 d} \right)}$$

for water at 20°C , surface tension $T = 72.7$ dynes/cm, and the density $d = 1.00$. Plot v against λ for water waves from $\lambda = 0$ to 30 cm, and graphically determine the group velocity for $\lambda = 1$ cm, 4 cm, and 25 cm.

19. Differentiate the equation for wave velocity of water waves in Prob. 18 and calculate the group velocity for $\lambda = 25$ cm, 50 cm, and 75 cm.

20. Two sources A and B emit waves of equal amplitude with phase angles of either 0° or 180° only. Show that, with a random distribution of these phases among the sources, the average total intensity will be two times that for any one of them alone. (NOTE: There are four phase combinations.)

21. Three sources A , B , and C emit waves of equal amplitude with phase angles of either 0° or 180° only. Show that with a random distribution of these phases among the sources the average total intensity will be three times that for any one of them alone. (NOTE: There are eight phase combinations.)

22. In Wiener's experiment a photographic film 5 cm long is placed in contact with a mirror and one edge then raised by inserting a strip of paper, 0.002 cm thick, between the two. Find the band separation to be found on the film if light of wavelength 5000 Å is used.

23. Write expressions for two simple periodic motions at right angles which when combined will produce resultant motions of the following types: (a) A linear motion of amplitude 3 and frequency 20 per sec, along a line making an angle of 60° with the positive x axis. (b) A circular motion of radius 3 and frequency 20 per sec, the center at the origin, and the initial position of the reference point at $+3$ on the y axis. (c) An elliptical motion with semiminor axis of 2 in the x direction, semimajor axis of 4 in the y direction, frequency of 20 per sec, and initial position of $+2$ on the x axis.

24. Write equations for two simple periodic motions at right angles [*i.e.*, $y = f_1(t)$, $z = f_2(t)$] that give (a) a linear motion of amplitude 5 and frequency 10 per sec along a line making an angle of 50° with the $+z$ axis, (b) the same as (a) but with an angle of 150° , and (c) a circular motion of radius 5, frequency 10 per sec, and center at the origin.

25. For the type of waves described in Prob. 18, find the exact value of the wavelength for which the wave and group velocities are equal, and determine this velocity.

CHAPTER 13

INTERFERENCE OF TWO BEAMS OF LIGHT

It was stated at the beginning of the last chapter that two beams of light may be made to cross each other without either one producing any modification of the other after it passes beyond the region of crossing. In this sense the two beams do not interfere with each other. However, in the region of crossing, where both beams are acting at once, we are led to expect from the considerations of the preceding chapter that the resultant amplitude and intensity may be very different from the sum of those contributed by the two beams acting separately. This modification of intensity obtained by the superposition of two or more beams of light we call *interference*. If the resultant intensity is zero or in general less than we expect from the separate intensities, we have *destructive* interference, while if it is greater, we have *constructive* interference. The phenomenon in its simpler aspects is rather difficult to observe, because of the very short wavelength of light, and therefore was not recognized as such in the time prior to 1800 when the corpuscular theory of light was predominant. The first man successfully to demonstrate the interference of light, and thus establish its wave character, was Thomas Young. In order to understand his crucial experiment performed in 1801, we must first consider the application to light of an important principle which holds for any type of wave motion.

13.1. Huygens' Principle. When waves pass through an aperture, or past the edge of an obstacle, they always spread to some extent into the region which is not directly exposed to the oncoming waves. This phenomenon is called *diffraction*. In order to explain this bending of light, Huygens nearly three centuries ago proposed the rule that *each point on a wave front may be regarded as a new source of waves*.^{*} This principle

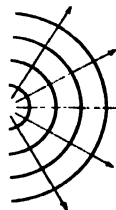


FIG. 13A. Diffraction of waves at a small aperture.

^{*} The "waves" envisioned by Huygens were not continuous trains but rather a series of random pulses. Furthermore, he supposed the secondary waves to be effective only at the point of tangency to their common envelope, thus denying the possibility of diffraction. The correct application of the principle was first made by Fresnel, more than a century later.

has very far-reaching applications and will be used later in discussing the diffraction of light, but we shall consider here only a very simple proof of its correctness. In Fig. 13A let a set of plane waves approach the barrier AB from the left, and let the barrier contain an opening S of width somewhat smaller than the wavelength. At all points except S the waves will be either reflected or absorbed, but S will be free to produce a disturbance behind the screen. It is found experimentally, in agreement with the above principle, that the waves spread out from S in the form of semicircles.

Huygens' principle as shown in Fig. 13A can be illustrated very successfully with water waves. An arc lamp on the floor, with a glass-bottomed tray or tank above it, will cast shadows of waves on a white ceiling. A vibrating strip of metal or a wire fastened to one prong of a tuning fork of low frequency will serve as a source of waves at one end of the tray. If an electrically driven tuning fork is used, the waves may be made apparently to stand still by placing a slotted disk on the shaft of a motor in front of the arc lamp. The disk is set rotating with the same frequency as the tuning fork to give the stroboscopic effect. The latter experiment can be performed for a fairly large audience and is well worth doing. Descriptions of diffraction experiments in light will be given in Chap. 15.

If the experiment in Fig. 13A be performed with light, one would naturally expect, from the fact that light generally travels in straight lines, that merely a narrow patch of light would appear at D . However, if the slit is made very narrow, an appreciable broadening of this patch is observed, its breadth increasing as the slit is further narrowed. This remarkable evidence that light does not always travel in straight lines was mentioned at the very beginning of this book (Sec. 1.1 and Fig. 1A). When the screen CE is replaced by a photographic plate, a picture like the one shown in Fig. 13B is obtained. The light is most intense in the forward direction, but its intensity decreases slowly as the angle increases. If the slit is small compared with the wavelength of light, the intensity does not come to zero even when the angle of observation becomes 90° (Sec. 15.3). While this brief introduction to Huygens' principle will be sufficient for an understanding of the interference phenomena we are to discuss, we shall return in Chaps. 15 and 18 to a more detailed consideration of diffraction at a single opening.

13.2. Young's Experiment. The original experiment performed by Young is shown schematically in Fig. 13C. Sunlight was first allowed

to pass through a pinhole S and then, at a considerable distance away, through two pinholes S_1 and S_2 . The two sets of spherical waves emerging from the two holes interfered with each other in such a way as to form a symmetrical pattern of varying intensity on the screen AC . Since this early experiment was performed, it has been found convenient to replace the pinholes by narrow slits and to use a source giving monochromatic light, *i.e.*, light of a single wavelength. In place of spherical wave fronts we now have cylindrical wave fronts, represented equally well in



FIG. 13B. Photograph of the diffraction of light from a slit of width 0.001 mm.

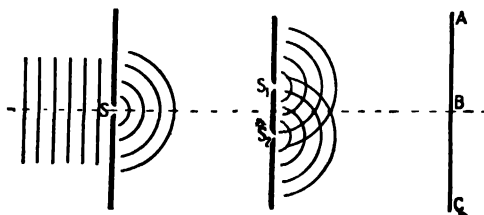


FIG. 13C. Experimental arrangement for Young's double-slit experiment.

two dimensions by the same Fig. 13C. If the circular lines represent crests of waves, the intersections of any two lines represent the arrival at those points of two waves with the same phase or with phases differing by a multiple of 2π . Such points are therefore those of maximum disturbance or brightness. A close examination of the light on the screen will reveal evenly spaced light and dark bands or fringes, similar to those shown in Fig. 13D. Such photographs are obtained by replacing the screen AC of Fig. 13C by a photographic plate.

A very simple demonstration of Young's experiment can be accomplished in the laboratory or lecture room by setting up a single-filament lamp L (Fig. 13E) at the front of the room. The straight vertical filament S acts as the source and first slit. Double slits for each observer

can be easily made from small photographic plates about 1 to 2 in. square. The slits are made in the photographic emulsion by drawing the point of a penknife across the plate, guided by a straight edge. The plates need not be developed or blackened but can be used as they are. The lamp is now viewed by holding the double slit D close to the eye E and looking at the lamp filament. If the slits are close together, *e.g.*, 0.2 mm apart, they give widely spaced fringes, whereas slits farther apart, *e.g.*, 1 mm, give very narrow fringes. A piece of red glass F , placed adjacent to and above another of green glass in front of the lamp, will show that the red waves produce wider fringes than the green, which we shall see is due to their greater wavelength.

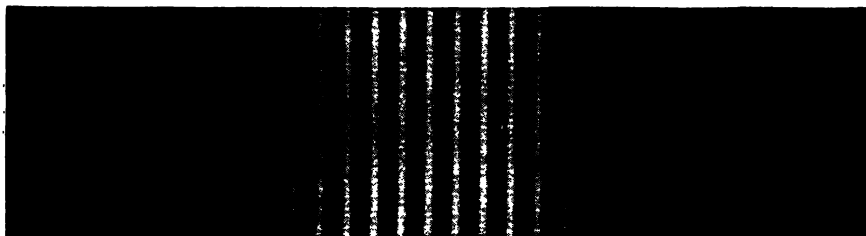


FIG. 13D. Interference fringes produced by a double slit using the arrangement shown in Fig. 13C.

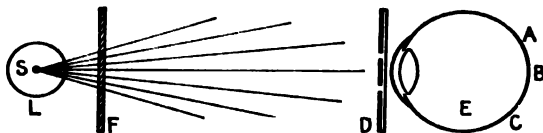


FIG. 13E. Simple method for observing interference fringes.

Frequently one wishes to perform accurate experiments by using more nearly monochromatic light than that obtained by white light and a red or green glass filter. Perhaps the most convenient method is to use the sodium arc now available on the market, or a d-c carbon arc arranged as follows: A small $\frac{1}{8}$ -in. hole, 1 in. deep, is drilled in the end of the positive carbon and filled with common salt. When the arc has run several minutes, the hole is refilled. After several refillings a very bright source of sodium light, almost entirely of wavelength 5893 Å, is obtained. Monochromatic green light can be obtained from any mercury arc by sending the light through a special glass filter now on the market. Such a filter transmits only the green line, $\lambda 5461$.

13.3. Interference Fringes from a Double Source. We shall now derive an equation for the intensity at any point P on the screen (Fig.

13*F*) and investigate the spacing of the interference fringes. Assuming that the source slit S is equidistant from S_1 and S_2 , the light vibrations at the two slits will be in the same phase at any instant, and we may represent either of them by the equation (Eq. 11*a*) of simple periodic motion,

$$y = r \sin 2\pi \frac{t}{T}$$

where r is the amplitude, T the period, and t the time.

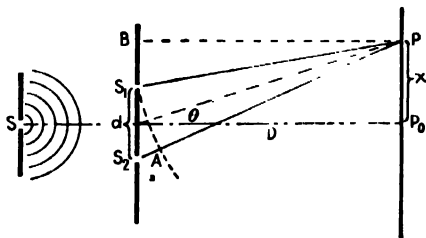


FIG. 13*F*. Schematic diagram of the optical paths of interfering beams in Young's experiment.

In the wave traveling from S_1 to P , the phase difference will have the value $2\pi S_1P/\lambda$ and hence at the point P will have the motion

$$y_1 = r_1 \sin 2\pi \left(\frac{t}{T} - \frac{S_1P}{\lambda} \right)$$

for the light from S_1 , and the motion

$$y_2 = r_2 \sin 2\pi \left(\frac{t}{T} - \frac{S_2P}{\lambda} \right)$$

for the light from S_2 . These two equations represent simple periodic motions of the same frequency, and we have exactly the problem treated in Sec. 12.1 by the principle of superposition. Our present equations have the form of Eqs. 12*a*, namely,

$$\begin{aligned} y_1 &= r_1 \sin (\omega t + \alpha_1) \\ y_2 &= r_2 \sin (\omega t + \alpha_2) \end{aligned}$$

where now $\omega = 2\pi/T$, $\alpha_1 = -2\pi S_1P/\lambda$, and $\alpha_2 = -2\pi S_2P/\lambda$. Therefore, from Eq. 12*f* we may write for the resultant motion

$$y = y_1 + y_2 = R \sin (\omega t + \theta)$$

in which R and θ are to be found from Eqs. 12*d* and 12*e*, respectively. The new phase constant θ is of no interest to us here, as we are primarily interested in the resultant intensity, which is proportioned to R^2 .

If, as is usually the case, the two slits S_1 and S_2 are of equal width and very close together, the amplitudes r_1 and r_2 will be so nearly equal that we may assume them to be so and put $r_1 = r_2 = r$. Then we have from Eq. 12*g*, for the resultant intensity,

$$I = R^2 = 4r^2 \cos^2 \frac{\delta}{2} \quad (13a)$$

where δ is the phase difference $\alpha_1 - \alpha_2$ between the two superimposed vibrations, given by the relation

$$\delta = \frac{2\pi}{\lambda} \cdot (\text{path difference}) = \frac{2\pi}{\lambda} (S_2P - S_1P) \quad (13b)$$

It now remains to find an expression for the path difference in terms of the distance x on the screen from the central point P_0 , the separation of the slits d , and the distance D from the slits to the screen. From the relations between the squares on the sides of a right triangle, we first write for the triangle BPS_2

$$(S_2P)^2 = D^2 + \left(x + \frac{d}{2}\right)^2$$

and for the triangle BPS_1

$$(S_1P)^2 = D^2 + \left(x - \frac{d}{2}\right)^2$$

Taking the difference between these two equations, D drops out, and

$$(S_2P)^2 - (S_1P)^2 = \left(x + \frac{d}{2}\right)^2 - \left(x - \frac{d}{2}\right)^2$$

which gives

$$(S_2P)^2 - (S_1P)^2 = 2xd$$

Factoring the left side of this equation,

$$(S_2P - S_1P)(S_2P + S_1P) = 2xd$$

or

$$S_2P - S_1P = \frac{2xd}{S_2P + S_1P}$$

In general, when Young's experiment is performed, D is some thousand times larger than d or x , so that $S_2P + S_1P$ may be replaced by $2D$

without altering the equality by more than a very small fraction of 1 per cent. We find

$$S_2P - S_1P = \frac{2xd}{2D} = \frac{xd}{D} \quad (13c)$$

This is the value of the path difference to be substituted in Eq. 13b to obtain the phase difference δ . Now Eq. 13a for the intensity has maximum values of $4r^2$ whenever δ is an integral multiple of 2π , and according to Eq. 13b this will occur when the path difference is an integral multiple of λ . Hence we have

$$\frac{xd}{D} = 0, \lambda, 2\lambda, 3\lambda, \dots = m\lambda$$

or

$$x = m\lambda \frac{D}{d} \quad \text{BRIGHT FRINGES} \quad (13d)$$

The minimum value of the intensity is zero, and as is seen from Eq. 13a this occurs when $\delta = \pi, 3\pi, 5\pi, \dots$. For these points

$$\frac{xd}{D} = \frac{\lambda}{2}, \frac{3\lambda}{2}, \frac{5\lambda}{2}, \dots = \left(m + \frac{1}{2}\right)\lambda$$

or

$$x = \left(m + \frac{1}{2}\right)\lambda \frac{D}{d} \quad \text{DARK FRINGES} \quad (13e)$$

The whole number m , which characterizes a particular bright fringe, is called the *order* of interference. Thus the fringes with $m = 0, 1, 2, \dots$ are called the zero, first, second, etc., orders.

The distance on the screen between two fringes of orders m and $m + 1$ may be obtained from Eq. 13d by taking the difference

$$x_{m+1} - x_m = (m + 1)\lambda \frac{D}{d} - \lambda \frac{D}{d} = \lambda \frac{D}{d} \quad (13f)$$

It is the same as the separation between dark fringes,

$$x_{m+\frac{1}{2}} - x_{m-\frac{1}{2}} = \left(m + \frac{1}{2}\right)\lambda \frac{D}{d} - \left(m - \frac{1}{2}\right)\lambda \frac{D}{d} = \lambda \frac{D}{d} \quad (13g)$$

According to these equations the spacing of the fringes is constant (independent of m), in agreement with the observed pattern of Fig. 13D. It is directly proportional to the slit-screen distance D , inversely proportional to the separation of slits d , and directly proportional to wavelength λ . A measurement of the spacing of the fringes thus gives us a direct determination of λ in terms of known quantities.

13.4. Intensity Distribution in the Fringe System. Let us now consider the physical reason for the formation of these dark and bright fringes. We have found that when the path difference $S_2P - S_1P$ is a whole number of wavelengths, the point P is the center of a bright fringe. For such a point the additional distance that one wave travels will be S_2A (Fig. 13F), provided the broken curve S_1A is the arc of a circle with P as a center. Thus it is

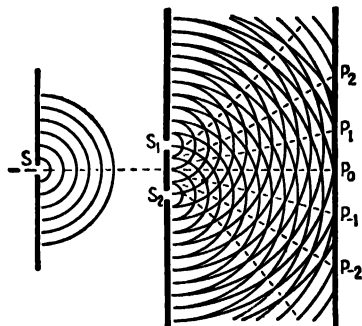


FIG. 13G. Wavelets from a double slit showing reinforcement along directions P_0, P_1, P_2 , etc.

clear that if S_2A contains a whole number of wavelengths, the two waves will reach P in the same phase, and the resultant amplitude will be twice that due to either wave alone. On the other hand, if S_2A is $(m + \frac{1}{2})\lambda$, i.e., an integral number of wavelengths plus an additional half wavelength, the waves reach P exactly in opposite phase, and the resultant amplitude will be zero. Figure 13G shows two complete sets of waves diverging from S_1 and S_2 , the semicircles representing the crests of the waves, a distance λ apart. The intensity will be a maximum wherever a crest falls on a crest. This will obviously occur at P_0 , which is equidistant from S_1 and S_2 , and also at P_1, P_{-1}, P_2, P_{-2} , etc., because each of these points is some whole number of wavelengths farther from one slit than from the other. If the screen is moved toward or away from the slits the spacing of these maxima decreases or increases nearly in direct proportion to the distance; i.e., the disturbance is a maximum at all points in space along the broken lines shown in the figure. These are not straight lines, as would be required by our simple equation 13d. They are actually hyperbolas, since the hyperbola is a curve for which the difference in the distance from two fixed points is a constant. However, when the wavelength is small and the distance to the screen large, the deviation from straight lines is small enough to be negligible for practical purposes.

To find the intensity on the screen at points between the maxima, we apply the vector method of compounding amplitudes described in Sec. 12.2 and illustrated for the present case in Fig. 13H. For the maxima, the angle δ is zero and the component amplitudes r_1 and r_2 are parallel, with the resultant $R = 2r$. For the minima, r_1 and r_2 are in opposite directions and $R = 0$. In general, for any value of δ , R is the

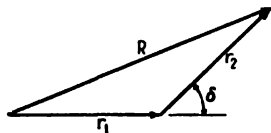


FIG. 13H. Illustrating the composition of two waves of the same frequency and amplitude but different phase.

closing side of the triangle. The value of R^2 , which measures the intensity, is then given by Eq. 13a and varies according to $\cos^2(\delta/2)$. In Fig. 13I the solid curve represents a plot of the intensity against the phase difference.

In concluding our discussion of these fringes, one question of fundamental importance should be considered. If the two beams of light arrive at a point on the screen exactly out of phase, they interfere destructively and the resultant intensity is zero. One may well ask what becomes of the *energy* of the two beams, since the law of conservation of energy tells us that energy cannot be destroyed. The answer to this question is that the energy which apparently disappears at the minima

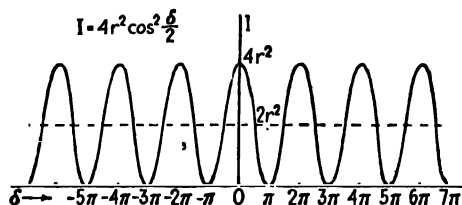


FIG. 13I. Intensity distribution for the interference fringes from two beams.

actually is still present at the maxima, where the intensity is greater than would be produced by the two beams acting separately. In other words, the energy is not destroyed but merely redistributed in the interference pattern. The *average* intensity on the screen is exactly that which would exist in the absence of interference. Thus, as shown in Fig. 13I, the intensity in the interference pattern varies between $4r^2$ and zero. Now each beam acting separately would contribute r^2 , and so without interference we would have a uniform intensity of $2r^2$, as indicated by the broken line. To obtain the average intensity on the screen for n fringes, we note that the average value of the square of the cosine is $\frac{1}{2}$. This gives, by Eq. 13a, $I = 2r^2$, justifying the statement made above, and it shows that no violation of the law of conservation of energy is involved in the interference phenomenon.

13.5. Fresnel's Biprism.* Soon after the double-slit experiment was performed by Young, the objection was raised that the bright fringes he observed were probably due to some complicated modification of the light by the edges of the slits and not to true interference. Thus the wave theory of light was still questioned. Not many years passed,

* Augustin Fresnel (1788–1827). Most notable French contributor to the theory of light. Trained as an engineer, he became interested in light, and in 1814–1815 he rediscovered Young's principle of interference and extended it to complicated cases of diffraction. His mathematical investigations gave the wave theory a sound foundation.

however, before Fresnel brought forward several new experiments in which the interference of two beams of light was proved in a manner not open to the above objection. The first of these is called the Fresnel biprism experiment.

A schematic diagram of the biprism experiment is shown in Fig. 13J. The thin double prism P refracts the light from the slit source S into

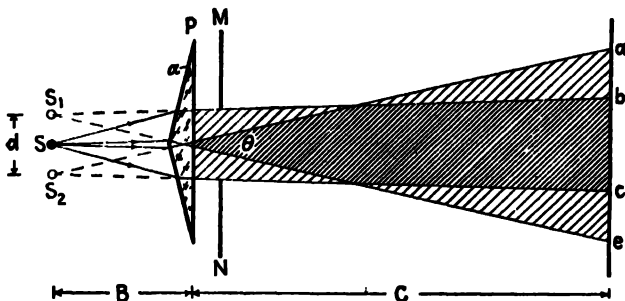


FIG. 13J. Diagram of Fresnel's biprism experiment.

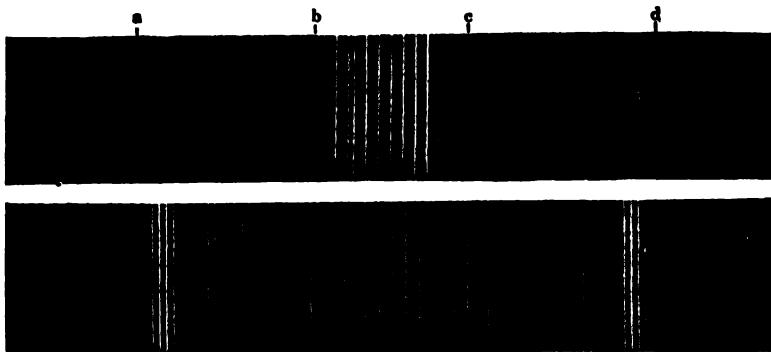


FIG. 13K. Interference and diffraction fringes produced in the Fresnel biprism experiment.

two overlapping beams ac and be . If screens M and N are placed as shown in the figure, interference fringes are observed only in the region bc . When the screen ae is replaced by a photographic plate, a picture like the upper one in Fig. 13K is obtained. The closely spaced fringes in the center of the photograph are due to interference, while the wide fringes at the edge of the pattern are due to diffraction. These wider bands are produced by the vertices of the two prisms, each of which acts as a straight edge, giving a pattern which will be discussed in detail in Chap. 18. When the screens M and N are removed from the light path, the two beams will overlap over the whole region ad . The lower photograph in Fig. 13K shows for this case the equally spaced interference

fringes superimposed on the diffraction pattern of a wide aperture. (For the diffraction pattern above, without the interference fringes, see Fig. 187.) With such an experiment Fresnel was able to produce interference without relying upon diffraction to bring the interfering beams together.

Just as in Young's double-slit experiment, the wavelength of light can be determined from measurements of the interference fringes produced by the biprism. Calling B and C the distances of the source and screen, respectively, from the prism P , d the distance between the virtual images S_1 and S_2 , and Δx the distance between the successive fringes on the screen, the wavelength of the light is given from Eq. 13f as

$$\lambda = \frac{\Delta x d}{B + C} \quad (13h)$$

Thus the virtual images S_1 and S_2 act as did the two slit sources in Young's experiment.

To find d , the separation of the virtual sources, we make use of the fact that for a prism of very small refracting angle the deviation of a ray is given by $(n - 1)\alpha$, n being the index of refraction of the prism and α its refracting angle. Hence, from Fig. 13J,

$$\frac{\theta}{2} = (n - 1)\alpha \quad (13i)$$

Now both α and n may be measured by placing the biprism on a spectrometer, so that θ can be found. Then $d = B\theta$, and we obtain for the wavelength

$$\lambda = \frac{B\theta}{B + C} \Delta x = \frac{2B(n - 1)\alpha}{B + C} \Delta x \quad (13j)$$

This method is rather laborious, and in practice it is much more convenient to measure the angle θ directly on the spectrometer. Parallel light from the collimator, when incident on both halves of the biprism, divides into two beams making an angle θ with each other, as does the central ray from S in Fig. 13J. The angular separation of the two slit images in the telescope is then equal to θ . A still simpler determination of θ may be made by holding the prism close to one eye and viewing a round frosted light bulb. At a certain distance from the light the two images may be brought to the point where their inner edges just touch. The diameter of the bulb divided by the distance from the bulb to the prism then gives θ directly.

Fresnel biprisms are easily made from a small piece of glass, such as half a microscope slide, by beveling about $\frac{1}{8}$ to $\frac{1}{4}$ in. on one side. This requires very little grinding with ordinary abrasive materials and polishing with rouge, since the angle required is only about 1° .

13.6. Fresnel's Mirrors. Another experiment illustrating the interference between two beams of light is known as the Fresnel mirror experi-

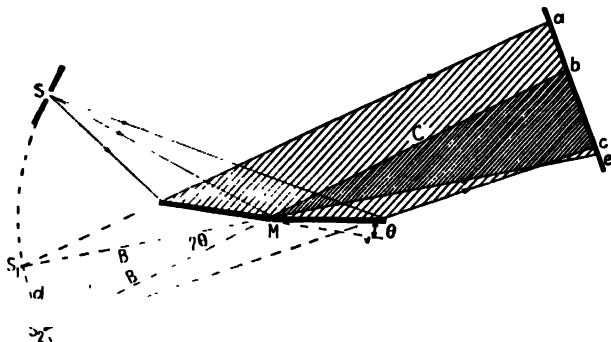


FIG. 13L. Diagram of the Fresnel double-mirror experiment.

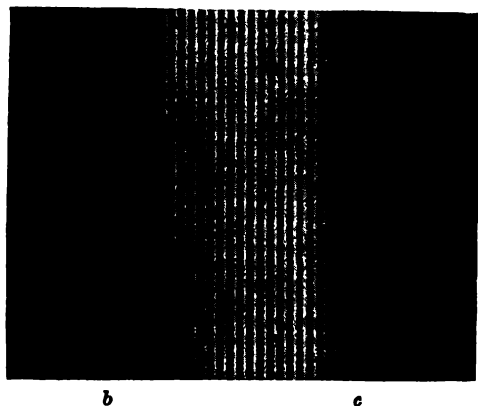


FIG. 13M. Interference fringes produced in the Fresnel double-mirror experiment.

ment. Light from a narrow slit S is split into two beams by reflection from two mirrors placed close together so that their planes make a small angle θ with each other, as illustrated in Fig. 13L. On that part of the screen where the two beams overlap, interference fringes are obtained of the type shown in Fig. 13M. The explanation of these fringes is similar to that for the double slit and biprism. After reflection from the mirrors, the light arriving at the screen appears to come from virtual sources

S_1 and S_2 . The relations between fringe separation and the geometry of the figure are such that, if we call d the distance between the virtual sources, C the distance Mb , and λ the wavelength, the distance Δx between fringes on the screen will be related to these quantities by Eq. 13h. The interval d can be measured directly by the second or third method described above for the biprism, or it can be calculated from the fact that $d = 2B\theta$ (see Fig. 13L), θ now being the angle between the mirrors. In terms of this angle, the wavelength is given by

$$\lambda = \frac{2B\theta}{B + C} \Delta x \quad (13k)$$

The Fresnel double-mirror experiment is usually performed on an optical bench with the light reflected from the mirrors at nearly grazing angles. Two pieces of ordinary plate glass about 2 in. square make a very good double mirror. One plate should have an adjusting screw for changing the angle θ , and the other a screw for making the two mirror edges parallel.

13.7. Coherent Sources. It will be noticed that the three successful methods of demonstrating interference we have discussed so far have one important feature in common: The two interfering beams are always derived from the same source of light. We find by experiment that it is impossible to obtain interference fringes from two separate sources, such as two lamp filaments set side by side. This failure is caused by the fact that the light from any one source is not an infinite train of waves. On the contrary, there are sudden changes in phase occurring in very short intervals of time (of the order of 10^{-8} sec). This point has already been mentioned in Sec. 11.3. Thus, although interference fringes may exist on the screen for such a short interval, they will shift their position each time there is a phase change, with the result that no fringes at all will be seen. In Young's experiment and in Fresnel's mirrors and biprism, the two sources S_1 and S_2 always have a point-to-point correspondence of phase, since they are both derived from the same source. If the phase of the light from a point in S_1 suddenly shifts, that of the light from the corresponding point in S_2 will shift simultaneously. The result is that the *difference* in phase between any pair of points in the two sources always remain constant, and so the interference fringes are stationary. It is a characteristic of any interference experiment with light that the sources must have this point-to-point phase relation, and sources that have this relation are called *coherent sources*.

If in Young's experiment the source slit S (Fig. 13C) is made too wide, or the angle between the rays which leave it too large, the double slit no longer represents two coherent sources, and the interference fringes disappear. This subject will be discussed in more detail at the end of Chap. 16, The Double Slit.

13.8. Lloyd's Mirror. The experiment known as Lloyd's mirror is important in any treatment of the nature of light, for it shows, in addition to the interference of two coherent beams of light, the phase change of light as it is reflected at grazing incidence from the surface of glass.

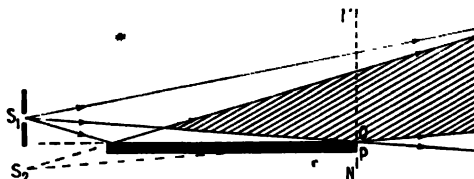
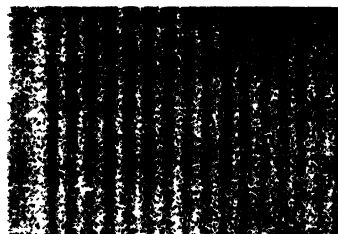


FIG. 13N. Diagram for the Lloyd's-mirror experiment.



(a) Taken with visible light $\lambda 4358 \text{ \AA}$. (After White.)



(b) Taken with X rays $\lambda 8.33 \text{ \AA}$. (After Kellstrom.)

FIG. 13O. Interference fringes produced with Lloyd's mirror.

Light from a narrow slit S_1 , in Fig. 13N, is incident at a grazing angle on the surface of a fairly long and flat strip of glass. The light is reflected from the glass in such a manner that its arrival at the screen is essentially the same as though it started from the virtual source S_2 . In addition to the reflected light arriving at the screen there is also the light coming directly from the source S_1 without reflection. In the region of overlapping of these two beams, interference occurs, and it can be observed as a system of fringes on the screen in the region bc .

An important feature of the Lloyd's-mirror experiment lies in the fact that when the screen is placed in contact with the end of the mirror (in the position MN , Fig. 13N), the edge O of the reflecting surface comes at the center of a *dark* fringe, instead of a bright one as might be expected. This means that one of the two beams has undergone a phase change of π .

Since the direct beam could not change phase, this experimental observation is interpreted to mean that the reflected light has changed phase at reflection. Two photographs of fringes formed by the Lloyd's-mirror experiment are reproduced in Fig. 13O, one taken with visible light and the other with X rays.

If the light from source S_1 in Fig. 13N is allowed to enter the end of the glass plate by moving the latter up, and to be internally reflected from the upper glass surface, fringes will again be observed in the interval OP , with a dark fringe at O . This again shows that there is a phase change of π at reflection. As will be shown in Chap. 28, this is not in contradiction with the discussion of phase change given in Sec. 11.8. In this instance the light is incident at an angle greater than the critical angle for total reflection.

The Lloyd's-mirror experiment is readily set up for demonstration purposes as follows: A carbon arc, followed by a colored glass filter and a narrow slit, serves as a source. A strip of ordinary plate glass 1 to 2 in. wide and a foot or more long makes an excellent mirror. A magnifying glass focused on the far end of the mirror enables one to observe the fringes shown in Fig. 13O. Internal fringes can be observed by polishing the ends of the mirror to allow the light to enter and leave the glass, and by roughening one of the glass faces with coarse emery.

13.9. Billet's Split Lens. Another device for producing interference fringes is known as Billet's split lens. In this experiment (Fig. 13P) half lenses are placed close together to form two real images S_1 and S_2 of the slit S . S_1 and S_2 now act in the same way as the double slit in Young's experiment. Fringes are observed in the overlapping region bc . An ordinary lens and a biplate, consisting of two identical plane-parallel plates inclined slightly to each other, will give the same result as a split lens.

13.10. Michelson* Interferometer. This is an instrument designed by Michelson in which light from an extended source is divided into two

* A. A. Michelson (1852-1937). American physicist of great genius. He early became interested in the velocity of light, and began experiments while an instructor in physics and chemistry at the Naval Academy, from which he graduated in 1873. It is related that the superintendent of the Academy asked young Michelson why he wasted his time on such useless experiments. Years later Michelson was awarded the Nobel prize (1907) for his work on light. Much of his work on the velocity of light (Sec. 19.5) was done during 10 years spent at the Case Institute of Technology. During the latter part of his life he was professor of physics at the University of Chicago, where many of his famous experiments on the interference of light were done.

parts by partial reflection. These beams are sent in quite different directions against plane mirrors, whence they are brought together again to form interference fringes. The arrangement is shown schematically in Fig. 13Q. The main optical parts consist of two highly polished plane mirrors M_1 and M_2 and two plane-parallel plates of glass G_1 and G_2 . Sometimes the rear side of the plate G_1 is lightly silvered (shown by the

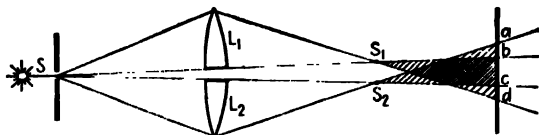


FIG. 13P. Diagram of Billet's split lens for producing interference fringes.

heavy line in the figure) so that the light coming from the source S is divided into (1) a reflected and (2) a transmitted beam of equal intensity. The light reflected normally from mirror M_1 passes through G_1 a third time and reaches the eye as shown. The light reflected from the mirror M_2 passes back through G_2 for the second time, is reflected from the surface of G_1 and into the eye. The purpose of the plate G_2 , called the compensating plate, is to render the path *in glass* of the two rays equal.

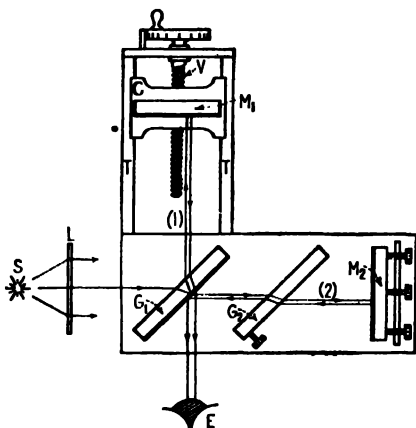


FIG. 13Q. Diagram of the Michelson interferometer.

This is not essential for producing fringes in monochromatic light, but it is indispensable when white light is used (Sec. 13.13). The mirror M_1 is mounted on a carriage C and can be moved along the well-machined ways or tracks T . This slow and accurately controlled motion is accomplished by means of the screw V which is calibrated to show the exact distance the mirror has been moved. To obtain fringes, the mirrors M_1 and M_2 are made exactly perpendicular to each other by means of screws shown on mirror M_2 .

Even when the above adjustments have been made, fringes will not be seen unless two important requirements are fulfilled. First, the light must originate from an *extended* source. A point source or a slit source, as used in the methods previously described, will not produce the desired system of fringes in this case. The reason for this will appear when we consider the origin of the fringes. Second, the light must in general be

monochromatic, or nearly so. Especially is this true if the distances of M_1 and M_2 from G_1 are appreciably different.

An extended source suitable for use with a Michelson interferometer may be obtained in any one of several ways. A sodium flame or a mercury arc, if large enough, may be used without the screen L shown in Fig. 13Q. If the source is small, a ground glass screen or a lens at L will extend the field of view. Looking at the mirror M_1 through the plate G_1 , one then sees the whole mirror filled with light. In order to obtain the fringes, the next step is to measure the distances of M_1 and M_2 to the back surface of G_1 roughly with a millimeter scale, and to move M_1 until they are the same to within a few millimeters. The mirror M_2 is now adjusted to be perpendicular to M_1 by observing the images of a common pin, or any sharp point, placed between the source and G_1 .

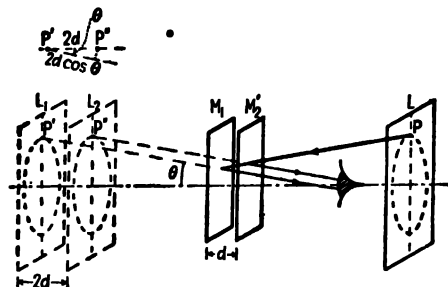


FIG. 13R. Diagram illustrating the formation of circular fringes in the Michelson interferometer.

Two pairs of images will be seen, one coming from reflection at the front surface of G_1 and the other from reflection at its back surface. When the tilting screws on M_2 are now turned until one pair of images falls exactly on the other, the interference fringes should appear. When they first appear, the fringes will not be clear unless the eye is focused on or near the back mirror M_1 , so the observer should look constantly at this mirror while searching for the fringes. When the fringes have been found, the adjusting screws are turned in such a way as to continually increase the width of the fringes, and finally a set of concentric circular fringes will be obtained. M_2 is then exactly perpendicular to M_1 , if the latter is at an angle of 45° with G_1 .

13.11. Circular Fringes. These are produced with monochromatic light when the mirrors are in exact adjustment, and are undoubtedly the most important type of fringes obtained with the Michelson interferometer. Their origin may be understood by reference to the diagram of

Fig. 13*R*. Here the real mirror M_2 has been replaced by its virtual image M'_2 formed by reflection in G_1 . M'_2 is then parallel to M_1 . Owing to the several reflections in the real interferometer, we may now think of the extended source as being behind the observer at L and forming two virtual images L_1 and L_2 in M_1 and M'_2 . These virtual sources are coherent in that the phases of corresponding points in the two are exactly the same at all instants. If d is the separation $M_1M'_2$, the virtual sources will be separated by $2d$. When d is exactly an integral number of half wavelengths, *i.e.*, the path difference $2d \cos \theta$ equal to an integral number of whole wavelengths, all rays of light reflected normal to the mirrors will be in phase. Rays of light reflected at an angle, however, will in general not be in phase. The path difference between the two rays coming to the eye from corresponding points P' and P'' is $2d \cos \theta$, as shown in the figure. The angle θ is necessarily the same for the two rays when M_1 is parallel to M'_2 so that the rays are parallel. Hence when the eye is focused to receive parallel rays (a small telescope is more satisfactory here, especially for large values of d) the rays will reinforce each other to produce maxima for those angles θ satisfying the relation*

$$2d \cos \theta = m\lambda \quad (13l)$$

Since for a given m , λ , and d the angle θ is constant, the maxima will lie in the form of circles about the foot of the perpendicular from the eye to the mirrors. By expanding the cosine, it can be shown from Eq. 13*l* that the radii of the rings are proportional to the square roots of integers, as in the case of Newton's rings (Sec. 14.4). The intensity distribution across the rings follows Eq. 13*a*, in which the phase difference δ is given by

$$\delta = \frac{2\pi}{\lambda} 2d \cos \theta$$

With monochromatic light the circular fringes are visible for very large path differences, the limit being set only by the fact that no actual source gives perfectly monochromatic light. If there is even a small range of wavelengths present in the light from the source, the fringes formed by the different components will be differently spaced and will mask all interference at sufficiently large values of d . Using the very nearly monochromatic light of the red cadmium line, the fringes remain visible up to path differences of about 50 cm, or $d = 25$ cm. A study of the change of clearness or "visibility" of the fringes† with increasing path

* Under these conditions *minima* may be observed (see discussion at the end of Sec. 13.13).

† The visibility of fringes is quantitatively defined as $V = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}$, where I_{\max} and I_{\min} are the intensities at the maxima and minima of the fringe pattern.

difference gives information about the sharpness of a spectral line used for the source of light. The larger the path difference over which the fringes remain clear, the more monochromatic the light, or the sharper the line. This was one of the first uses to which Michelson put his interferometer. Maxima and minima in the visibility curve indicate that the line has a fine structure of two or more components. Thus it is found that with light of the sodium D lines the fringes will become alternately sharp and diffuse, as the fringes formed by the two lines get in and out of step. The number of fringes between two positions of maximum sharpness is about 1000, indicating that the wavelengths of the D lines differ by about 1 part in 1000. Since Michelson's method of inferring the structure of lines has been superseded by a more direct method (Sec. 14.6), it will not be described in detail here.

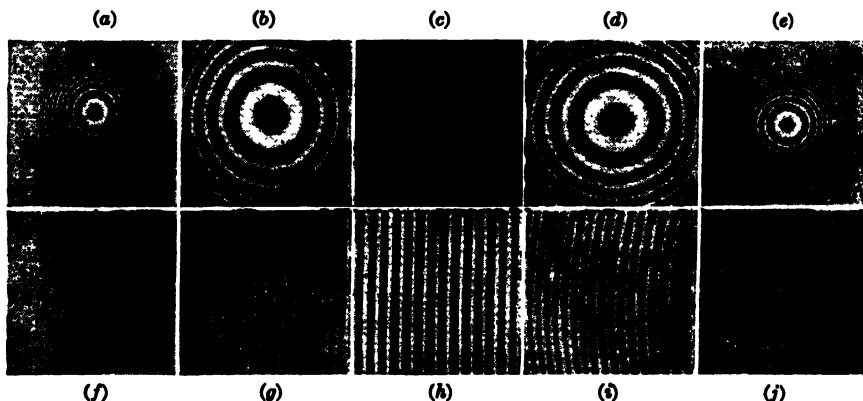


FIG. 13S. Appearance of the various types of fringes observed in the Michelson interferometer.

Starting with M_1 a few centimeters beyond M'_2 , the fringe system will have the general appearance shown in (a) of Fig. 13S, with the rings very closely spaced. If M_1 is now moved slowly toward M'_2 so that d is decreased, Eq. 13l shows that a given ring, characterized by a given value of the order m , must decrease its radius because the product $2d \cos \theta$ must remain constant. The rings therefore shrink and vanish at the center, a ring disappearing each time $2d$ decreases by λ , or d by $\lambda/2$. This follows from the fact that at the center $\cos \theta = 1$, so that Eq. 13l becomes

$$2d = m\lambda \quad (13m)$$

To change m by unity, d must change by $\lambda/2$. Now as M_1 approaches M'_2 the rings become more widely spaced, as indicated in Fig. 13S(b), until finally we reach a critical position where the central fringe has

spread out to cover the whole field of view, as shown in (c). This happens when M_1 and M'_2 are exactly coincident, for it is clear that under these conditions the path difference is zero for all angles of incidence. If the mirror is moved still farther, it effectively passes through M'_2 , and new widely spaced fringes appear, growing out from the center. These will gradually become more closely spaced as the path difference increases, as indicated in (d) and (e) of the figure.

13.12. Localized Fringes. If the mirrors M'_2 and M_1 are not exactly parallel, fringes will still be seen with monochromatic light for path differences not exceeding a few millimeters. In this case the space between the mirrors is wedge-shaped, as indicated in Fig. 13T. The two rays* reaching the eye from a point P on the source are now no longer parallel, but appear to diverge from a point P' near the mirrors. Thus to see these fringes clearly, the eye must be focused on or near the

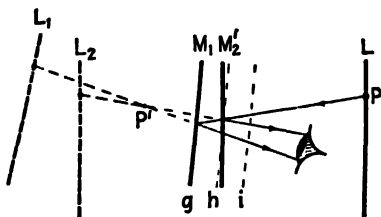


FIG. 13T. Diagram illustrating the formation of fringes with inclined mirrors in the Michelson interferometer.

rear mirror M_1 . The localized fringes are practically straight, because the variation of the path difference across the field of view is now due primarily to the variation of the thickness of the "air film" between the mirrors. With a wedge-shaped film, the locus of points of equal thickness is a straight line parallel to the edge of the wedge. The fringes are not exactly straight, however, if d has an appreciable value, because there is also some variation of the path difference with angle. They are in general curved and are always convex toward the thin edge of the wedge. Thus, with a certain value of d , we might observe fringes shaped like those of Fig. 13S(g). Decreasing d , they move to the left across the field, a new fringe crossing the center of the field each time d changes by $\lambda/2$. As we approach zero path difference, the fringes become straighter, until the point is reached where M_1 actually intersects M'_2 , when they are perfectly straight, as in (h). Beyond this point, they begin to curve in the opposite direction (i). The blank fields (f) and (j) indicate that this type of fringe cannot be observed for large path differences.

13.13. White-light Fringes. If a source of white light is used, no fringes will be seen at all except for a path difference so small that it does not exceed a few wavelengths. In observing these fringes, the mirrors are tilted slightly as for localized fringes, and the position of M_1

* When the term "ray" is used, here and elsewhere in discussing interference phenomena, it merely indicates the direction of the perpendicular to a wave front, and is in no way to suggest an infinitesimally narrow pencil of light.

is found where it intersects M'_2 . With white light there will then be observed a central dark fringe, bordered on either side by 8 or 10 colored fringes. This position is often rather troublesome to find using white light only. It is best located approximately beforehand by finding the place where the localized fringes in monochromatic light become straight. Then a very *slow* motion of M_1 through this region, using white light, will bring these fringes into view.

The fact that only a few fringes are observed with white light is easily accounted for when we remember that such light contains all wavelengths between 4000 and 7500 Å. The fringes for a given color are more widely spaced the greater the wavelength. Thus the fringes in different colors will only coincide for $d = 0$, as indicated in Fig. 13U. The solid curve represents the intensity distribution in the fringes for green light, and the broken curve that for red light. Clearly, only the central fringe will be uncolored, and the fringes of different colors will begin to separate at once on either side, producing various impure colors which are not the

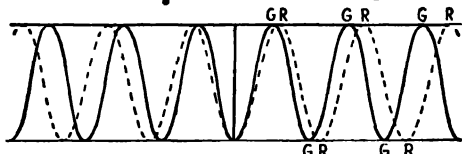


FIG. 13U. Illustrating the origin of white-light fringes with a dark fringe at the center.

saturated spectral colors. After 8 or 10 fringes, so many colors are present at a given point that the resultant color is essentially white. Interference is still occurring in this region, however, because a spectroscope will show a continuous spectrum with dark bands at those wavelengths for which the condition for destructive interference is fulfilled. White-light fringes are also observed in all the other methods of producing interference described above, if white light is substituted for monochromatic light. They are particularly important in the Michelson interferometer, where they may be used to locate the position of zero path difference, as we shall see in Sec. 13.14.

An excellent reproduction in color of these white-light fringes will be found in Michelson, "Light Waves and Their Uses," Plate II. The fringes in three different colors are also shown separately, and a study of these in connection with the white-light fringes is instructive as showing the origin of the various impure colors in the latter.

It was stated above that the central fringe in the white-light system, *i.e.*, that corresponding to zero path difference, is black when observed

in the Michelson interferometer. One would ordinarily expect this fringe to be white, since the two beams should be in phase with each other for any wavelength at this point, and in fact this is the case in the fringes formed with the other arrangements, such as the biprism. In the present case, however, it will be seen by referring to Fig. 13Q that while ray (1) undergoes an internal reflection in the plate G_1 , ray (2) undergoes an external reflection, with a consequent change of phase (Sec. 11.8). Hence the central fringe is black, if the black surface of G_1 is unsilvered. If it is silvered, the conditions are different and the central fringe may be white.

13.14. Applications of the Michelson Interferometer. The principal advantage of this form of interferometer over the earlier arrangements for producing interference lies in the fact that the two beams are here widely separated, and the path difference between them can be varied at will by moving the mirror M_1 or by introducing a refracting material in the path of one of the beams. Corresponding to these two ways of varying the path difference, there are two types of measurement which can be made with this interferometer. The first is the accurate measurement of distance in terms of the wavelength of light, which we shall discuss in this section. The second is the determination of indices of refraction, which will be briefly referred to at the beginning of Sec. 13.16.

When the mirror M_1 is moved slowly from one position to another, counting the number of fringes in monochromatic light which cross the center of the field of view will give a measure of the distance the mirror has moved in terms of λ , since by Eq. 13m we have, for the position d_1 corresponding to the bright fringe of order m_1 ,

$$2d_1 = m_1\lambda$$

and for d_2 , giving a bright fringe of order m_2 ,

$$2d_2 = m_2\lambda$$

Subtracting these two equations, we find

$$d_1 - d_2 = (m_1 - m_2) \frac{\lambda}{2} \quad (13n)$$

Hence the distance moved equals the number of fringes counted, multiplied by a half wavelength. Of course, the distance measured need not correspond to an integral number of half wavelengths. Fractional parts of a whole fringe displacement can easily be estimated to one-tenth of a fringe, and, with care, to one-fiftieth. The latter figure then gives the distance to an accuracy of $\frac{1}{50} \lambda$, or 5×10^{-7} cm for green light.

A small Michelson interferometer in which a microscope is attached to the moving carriage carrying M_1 is frequently used in the laboratory for measuring the wavelength of light. The microscope is focused on a fine glass scale, and the number of fringes, $m_1 - m_2$, crossing the mirror between two readings d_1 and d_2 on the scale gives λ , by Eq. 13n. The bending of a beam, or even of a brick wall, under pressure from the hand can be made visible and measured by attaching M_1 directly to the beam or wall.

The most important measurement made with the interferometer was the comparison of the standard meter in Paris with the wavelengths of intense red, green, and blue lines of cadmium by Michelson and Benoit. Clearly it would be impossible to count directly the number of fringes for a displacement of the movable mirror from one end of the standard meter to the other. Instead, nine intermediate standards (etalons) were used, of the form shown in Fig. 13V, each approximately twice the length of the other. The two shortest etalons were first mounted in an interferometer of special design (Fig. 13W), with a field of view covering the four mirrors, M_1 , M_2 , M'_1 , and M'_2 . With the aid of the white light

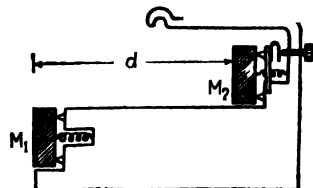


FIG. 13V. Diagram of one of the nine etalons used by Michelson.

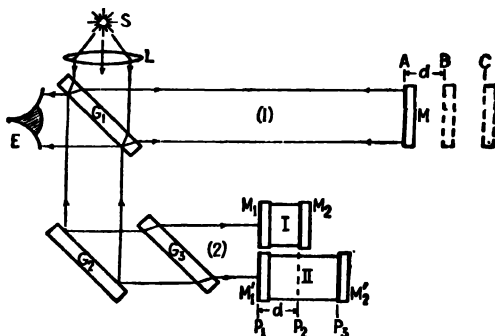


FIG. 13W. Diagram of the special Michelson interferometer used in accurately comparing the wavelength of light with the standard meter.

fringes the distances of M , M_1 , and M'_1 from the eye were made equal, as shown in the figure. Substituting the light of one of the cadmium lines for white light, M was then moved slowly from A to B , counting the number of fringes passing the cross hair. The count was continued until M reached the position B , which was exactly coplanar with M_2 , as judged by the appearance of the white-light fringes in the upper mirror

of the shorter etalon. The fraction of a cadmium fringe in excess of an integral number required to reach this position was determined, giving the distance M_1M_2 in terms of wavelengths. The shorter etalon was then moved through its own length, without counting fringes, until the white-light fringes reappeared in M_1 . Finally M was moved to C , when the white-light fringes appeared in M'_2 as well as in M_2 . The additional displacement necessary to make M coplanar with M'_2 was measured in terms of cadmium fringes, thus giving the exact number of wavelengths in the longer etalon. This was in turn compared with the length of a third etalon of approximately twice the length of the second, by the same process.

The longest intermediate standard was about 10 cm in length. This was compared with the prototype meter as shown in Fig. 13X. Starting

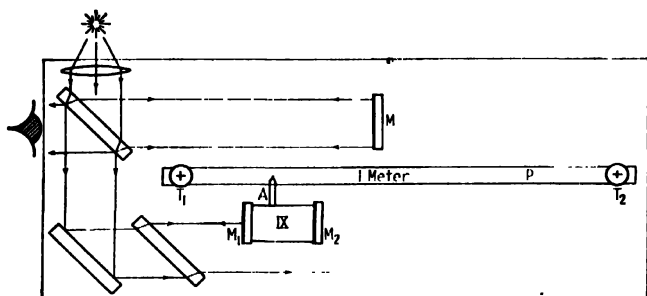


FIG. 13X. Comparison of the longest etalon with the standard meter.

with the pointer A in coincidence with the end mark under the microscope T_1 and M coplanar with M_2 as observed with white-light fringes, the etalon is moved through its own length until the fringes are centered in M_1 . M is then moved until they reappear in M_2 , and the etalon moved again, repeating the process until after nine displacements the pointer A appears in T_2 . The number of cadmium fringes required to make A coincide with the second end mark is finally determined.

It is important to notice that the error in the intercomparison of etalons is not cumulative. Thus the fractional part of a fringe measured in comparing twice the length of the first etalon with that of the second is only used to make sure of the *whole number* of the fringe nearest the cross hair when M is coplanar with M'_2 . The final stepping-off process with the longest etalon does involve an accumulated error, but this is at most much smaller than that made in setting on the end marks with the microscope.

The final results were, for the three cadmium lines:

Red line.....	1 m = 1,553,163.5λ or λ = 6438.4722 Å
Green line.....	1 m = 1,966,249.7λ or λ = 5085.8240 Å
Blue line.....	1 m = 2,083,372.1λ or λ = 4799.9107 Å

Not only has the standard meter been determined in terms of what we now believe to be an invariable unit, the wavelength of light, but we have also obtained absolute determinations of the wavelength of three spectrum lines, the red line of which is at present the primary standard in spectroscopy. More recent measurements on the red cadmium line have been made (see Sec. 14.7). It now is internationally agreed that in dry atmospheric air at 15°C and a pressure of 760 mm Hg the red cadmium line, produced under the conditions described by Michelson, has the wavelength

$$\lambda_r = 6438.4696 \text{ Å}$$

13.15. Twyman and Green's Interferometer. If a Michelson interferometer is illuminated with strictly parallel monochromatic light, produced by a point source at the principal focus of a well-corrected lens, it becomes a very powerful instrument for testing the perfection of optical parts such as prisms and lenses. The piece to be tested is placed in one of the light beams, and the mirror behind it is so chosen that the reflected waves, after traversing the test piece a second time, again become plane. These waves are then brought to interference with the plane waves from the other arm of the interferometer by another lens, at the focus of which the eye is placed. If the prism or lens is optically perfect, so that the returning waves are strictly plane, the field will appear uniformly illuminated. Any local variation of the optical path will, however, produce fringes in the corresponding part of the field, which are essentially the "contour lines" of the distorted wave front. Even though the surfaces of the test piece may be accurately made, the glass may contain regions that are slightly more or less dense. With the Twyman and Green interferometer these may be detected, and corrected for by local polishing of the surface.

13.16. Determination of Index of Refraction by Interference Methods. If a thickness t of a substance having an index of refraction n is introduced into the path of one of the interfering beams in the interferometer, the optical path in this beam is increased because of the fact that light travels more slowly in the substance and consequently has a shorter wavelength. The optical path (Eq. 11p) is now nt through the medium, whereas it was practically t through the corresponding thickness of air ($n = 1$). Thus the increase in optical path due to insertion of the sub-

stance is $(n - 1)t$.^{*} This will introduce $(n - 1)t/\lambda$ extra waves in the path of one beam, so if we call Δm the number of fringes by which the fringe system is displaced when the substance is placed in the beam, we have

$$(n - 1)t = (\Delta m)\lambda \quad (130)$$

In principle a measurement of Δm , t , and λ thus gives a determination of n .

In practice, the insertion of a plate of glass in one of the beams produces a discontinuous shift of the fringes so that the number Δm cannot be counted. With monochromatic fringes it is impossible to tell which fringe in the displaced set corresponds to one in the original set. With white light, the displacement in the fringes of different colors is very different because of the variation of n with wavelength, and the fringes disappear entirely. This illustrates the necessity of the compensating plate G_2 in Michelson's interferometer if white-light fringes are to be observed. If the plate of glass is very thin, these fringes may still be visible, and this affords a method of measuring n for very thin films. For thicker pieces, a practicable method is to use two plates of identical thickness, one in each beam, and to turn one gradually about a vertical axis, counting the number of monochromatic fringes for a given angle of rotation. This angle then corresponds to a certain known increase in effective thickness.

For the measurement of the index of refraction of gases, which can be introduced gradually into the light path by allowing the gas to flow into an evacuated tube, the interference method is the most practicable one. Several forms of refractometers have been devised especially for this purpose, of which we shall describe two, the Jamin refractometer and the Rayleigh refractometer.

Jamin's refractometer is shown schematically in Fig. 13Y. Monochromatic light from a broad source S is broken into two parallel beams (1) and (2) by reflection at the two parallel faces of a thick plate of glass G_1 . These two rays pass through to another identical plate of glass G_2 to recombine after reflection, forming interference fringes known as Brewster's fringes (see Sec. 14.7). If now the plates are parallel, the light paths will be identical. Suppose as an experiment we wish to measure the index of refraction of a certain gas at different temperatures and pressures. Two similar evacuated tubes T_1 and T_2 of equal length are placed in the two parallel beams. Gas is slowly admitted to tube T_2 . Counting the number of fringes Δm crossing the field while the gas

^{*} In the Michelson interferometer, where the beam traverses the substance twice in its back-and-forth path, t is twice the actual thickness.

reaches the desired pressure and temperature, the value of n can be found by applying Eq. 13o. It is found experimentally that at a given temperature the value $(n - 1)$ is directly proportional to the pressure. This is a special case of a theoretical law known as the *Lorentz-Lorenz* law* according to which

$$\frac{n^2 - 1}{n^2 + 2} = (n - 1) \frac{(n + 1)}{(n^2 + 2)} = \text{const.} \times \rho \quad (13p)$$

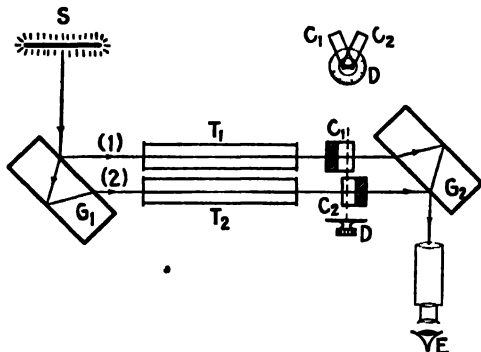


FIG. 13Y. Diagram of the Jamin refractometer.

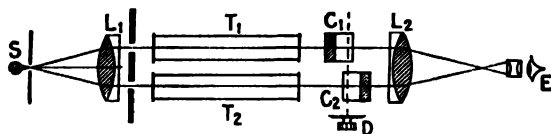


FIG. 13Z. Diagram of the Rayleigh refractometer.

Here ρ is the density of the gas. When n is very nearly unity, the factor $(n + 1)/(n^2 + 2)$ is nearly constant, as required by the above experimental observation.

In Rayleigh's† refractometer (Fig. 13Z) monochromatic light from a linear source S is made parallel by a lens L_1 , split into two beams by a

* H. A. Lorentz (1853–1928). For many years professor of mathematical physics at the University of Leyden, Holland. Awarded the Nobel prize (1902) for his work on the relations between light, magnetism, and matter, he also contributed notably to other fields of physics. Gifted with a charming personality and kindly disposition, he traveled a great deal, and was widely known and liked. By a strange coincidence L. Lorenz of Copenhagen derived the above law from the elastic-solid theory only a few months before Lorentz obtained it from the electromagnetic theory.

† Lord Rayleigh (third Baron) (1842–1919). Professor of physics at Cambridge University and the Royal Institution of Great Britain. Gifted with great mathematical ability and physical insight, he made important contributions to many fields of physics. His work on sound and on the scattering of light (Sec. 22.9) are the best known. He was a Nobel prize winner in 1904.

fairly wide double slit and sent through similar tubes and two compensating plates to be brought together again by lens L_2 to interfere in front of the observer.

The purpose of the compensating plates C_1 and C_2 in each of the above refractometers is to speed up the measurement and determination of the refractive index. As the two plates of equal thickness are rotated together by the single knob and dial D , one light path is shortened and the other lengthened. The device can therefore compensate for the path difference in the two tubes. The dial, if previously calibrated by counting fringes, can be made to read directly the index of refraction. The sensitivity of this device can be varied at will, a high sensitivity being obtained when the angle between the two plates is small and a low sensitivity when the angle is large.

Problems

1. Red light of wavelength 6800 \AA from a narrow slit falls on a double slit of separation (between centers) of $d = 0.026 \text{ cm}$. If the interference pattern is formed on a screen 100 cm away, what will be the linear separation between fringes on the screen?

2. Under the conditions of Prob. 1, what is the sign and magnitude of the percentage error in the distance of the tenth fringe from the center one, resulting from the approximation mentioned in the text above Eq. 13c?

3. Green light of wavelength 5120 \AA from a narrow slit is incident on a double slit of separation $d = 0.35 \text{ mm}$. Plot a curve giving the fringe separation as a function of the distance from the double slit.

4. Yellow light of wavelength 5800 \AA from a narrow slit is incident on a double slit. If the over-all separation of 10 fringes on a screen 160 cm away is 1.2 cm , find the double slit separation.

5. White light falling on a double slit of separation 1.5 mm forms colored fringes on a screen 100 cm away. If a pinhole is located in this screen at a distance of 2 mm from the central white fringe, what wavelengths within the visible spectrum will be absent from the transmitted light?

6. Solve Prob. 5 if the pinhole is located 4 mm from the central white fringe.

7. Interference fringes formed on a screen 80 cm from a double slit of separation 0.52 mm are measured to be 0.8 mm apart. Find the wavelength of the light and give its color.

8. A Fresnel biprism with refracting angles of 1° and index 1.53 is used to form interference fringes. Find the fringe separation for green light, $\lambda 5000$, when the distance between the source and the prism is 30 cm and the distance between the prism and the screen 70 cm .

9. Solve Prob. 8 if the distances 30 cm and 70 cm are interchanged.

10. Interference fringes of yellow light, $\lambda 5800$, are formed by Billet's split lens (see Fig. 13P). The distance from the source S to the lens L is 25 cm . The focal length of lens is 15 cm . The lens halves are separated 0.08 mm and the source-to-screen distance is 200 cm . Find the fringe separation.

11. Solve Prob. 10 if the distance S to L is 20 cm and other dimensions remain unchanged.

12. In moving one mirror M_1 of Michelson's interferometer a distance of 0.3220 mm, 1204 fringes are counted. Calculate the wavelength of light.

13. Solve Prob. 12 if 368 fringes are counted in moving the mirror 0.1220 mm.

14. Calculate the number of fringes that must be counted for green cadmium light for the shortest etalon used with Michelson's special interferometer, which had a length of 0.390 mm.

15. Solve Prob. 14 for red cadmium light.

16. Solve Prob. 14 for blue cadmium light.

17. The two tubes of a Jamin refractometer are 25.0 cm long. One contains a gas at a pressure of 10 cm Hg and the other is evacuated. If on removing the gas a shift of 20 fringes of green light 5760 Å is counted, what is the index of refraction of the gas at atmospheric pressure?

18. Solve Prob. 17 if 16 fringes are counted with blue light, $\lambda = 4500$ Å.

19. From Eq. 13/ prove that the radii of the circular fringes in the Michelson interferometer are proportional to the square roots of whole numbers.

20. If the path difference between the mirrors of a Michelson interferometer is 5 mm, what will be the angular radius of the fifth bright fringe in the circular pattern of fringes observed with green light $\lambda = 5000$ Å? (NOTE: Orders of interference m decrease from center of ring pattern outward.)

21. A source of microwaves, $\lambda = 1$ cm, is located at one end of a table 2 m long and 2 cm above the flat metal table top. Interference fringes are located at the far end of the table with a crystal detector. Determine the positions of the first three principal maxima, measured along a line at and perpendicular to the far end.

22. A pair of Fresnel mirrors making an angle of 1° with each other are located 1 m from a slit source emitting light of wavelength 6000 Å. Calculate the fringe separation on a screen 2 m beyond the intersection of the Fresnel mirrors.

23. A thin film of plastic of index $n = 1.45$ for light of wavelength 5890 Å is inserted in one arm of a Michelson interferometer. If a shift of 6.5 fringes is observed, find the film thickness.

24. The two compensating plates of a Jamin refractometer are inclined at a fixed angle of 5° with each other. One plate is vertical when fringes are first observed. Through what angle should they be rotated to produce a shift of 20 fringes of green light, 5500 Å, if the refractive index is $n = 1.500$? Assume plate thicknesses of 5 mm.

25. Two sources of sound having a pitch of 225 cycles per sec, and vibrating in phase, are separated by a distance of 20 ft. The velocity of sound is 1100 ft/sec. (a) Draw a sketch showing approximately the location of points of maximum intensity. How far apart, along the line joining the sources, are the points of maximum intensity? (b) The frequency of one of the sources is now increased to 226 cycles per sec. What now happens to the intensity pattern? At what rate does it sweep by a stationary observer situated on the line joining the sources?

CHAPTER 14

INTERFERENCE INVOLVING MULTIPLE REFLECTIONS

Some of the most beautiful effects of interference result from the multiple reflection of light between the two surfaces of a thin film of transparent material. These effects require no special apparatus for their production or observation and are familiar to anyone who has noticed the colors shown by thin films of oil on water, by soap bubbles, or by cracks in a piece of glass. We begin our investigation of this class of interference by considering the somewhat idealized case of reflection from a film with perfectly plane sides which are parallel to each other.

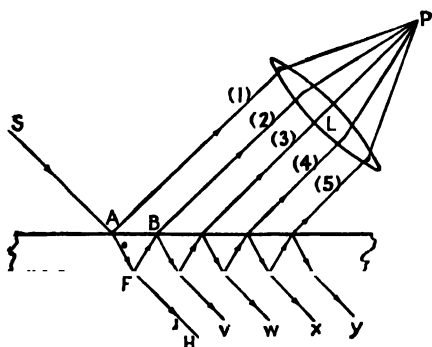


FIG. 14A. Multiple reflection of light between the boundaries of a thin film.

of the film. In each of these sets of course, the intensity decreases rapidly from one ray to the next. If the set of parallel reflected rays is now collected by a lens and focused at the point P , each ray will have traveled a different distance, and the phase relations may be such as to produce destructive or constructive interference at that point. In order to find the phase difference between these rays, we must first evaluate the difference in the optical path traversed by a pair of successive rays, such as rays (1) and (2).

In Fig. 14B let d be the thickness of the film, n its index of refraction, and λ the wavelength of the light, and let ϕ and ϕ' be the angles of incidence and refraction. If BD is perpendicular to ray (1), the optical paths from D and B to the focus of the lens will be equal. Therefore

14.1. Reflection from a Plane-parallel Film. Let a ray of light from a source S be incident on the surface of such a film at A (Fig. 14A). Part of this will be reflected as ray (1) and part refracted in the direction AF . Upon arrival at F , part of the latter will be reflected to B and part refracted toward H . At B the ray FB will be again divided. A continuation of this process yields two sets of parallel rays, one on each side

starting at A , ray (2) has the path AFB in the film and ray (1) the path AD in air. The difference in these optical paths (Eq. 11p) is given by

$$\text{Path difference } \Delta = n(afb) - AD$$

If BF is extended to intersect the perpendicular line AE at G , $AF = GF$ because of the equality of the angles of incidence and reflection at the lower surface. Thus we have

$$\Delta = n(GB) - AD = n(GC + CB) - AD$$

Now, since AC is drawn perpendicular to FB , the broken lines AC and DB represent two successive positions of a wave front reflected from the lower surface. The optical paths, according to the theorem of Malus (Sec. 1.9), must be the same by any ray drawn between two wave fronts, so we may write

$$n(CB) = AD$$

The path difference then reduces to

$$\Delta = n(GC) = n(2d \cos \phi') \quad (14a)$$

If this path difference is a whole number of wavelengths, we might expect rays (1) and (2) to arrive at the focus of the lens in phase with each other and produce a maximum of intensity. However, we must take account of the fact that ray (1) undergoes a phase change of π at reflection, while ray (2) does not, since it is internally reflected (Sec. 11.8). The condition

$$2nd \cos \phi' = m\lambda \quad \text{MINIMA} \quad (14b)$$

then becomes a condition for *destructive* interference as far as rays (1) and (2) are concerned.

Next we examine the phases of the remaining rays, (3), (4), (5), Since the geometry is the same, the path difference between rays (3) and (2) will also be given by Eq. 14a, but here there are only internal reflections involved, so that if Eq. 14b is fulfilled, ray (3) will be in the same phase as ray (2). The same holds for all succeeding pairs, and so we conclude that under these conditions rays (1) and (2) will be out of phase, but rays (2), (3), (4), . . . , will be in phase with each other. On the other hand, if conditions are such that

$$2nd \cos \phi' = (m + \frac{1}{2})\lambda \quad \text{MAXIMA} \quad (14c)$$

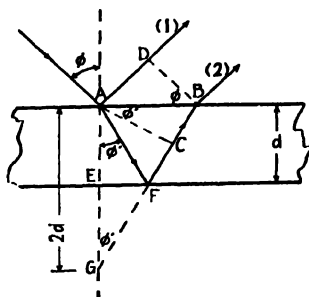


FIG. 14B. Optical path difference between two consecutive rays in multiple reflection (see Fig. 14A).

ray (2) will be in phase with (1), but (3), (5), (7), . . . will be out of phase with (2), (4), (6), Since (2) is more intense than (3), (4) more intense than (5), etc., these pairs cannot cancel each other, and since the stronger series combines with (1), the strongest of all, there will be a maximum of intensity.

For the minima of intensity, ray (2) is out of phase with ray (1), but (1) has a considerably greater amplitude than (2), so that these two will not completely annul each other. We can now prove that the addition of (3), (4), (5), . . . , which are all in phase with (2), will give just sufficient intensity to make up the difference and to produce complete darkness at the minima. Using a for the amplitude of the incident

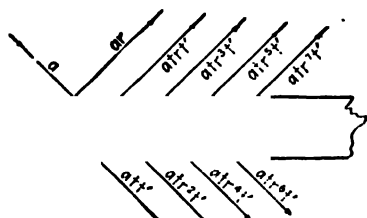


FIG. 14C. Showing the amplitude factors of successive rays in multiple reflection.

wave, r for the fraction of this reflected, and t or t' for the fraction transmitted in going from rare to dense or dense to rare, as was done in Stokes' treatment of reflection in Sec. 11.8, Fig. 14C is constructed and the amplitudes labeled as shown. In accordance with Eq. 11z, we have taken the fraction reflected internally and externally to be the same. Adding the amplitudes of

all the reflected rays but the first on the upper side of the film, we obtain the resultant amplitude,

$$\begin{aligned} R &= atrt' + atr^2t' + atr^3t' + atr^4t' + \cdots \\ &= atrt'(1 + r^2 + r^4 + r^6 + \cdots) \end{aligned} \quad (14d)$$

Since r is necessarily less than 1, the geometrical series in parentheses has a finite sum equal to

$$\frac{1}{1 - r^2}$$

Substituting in Eq. 14d,

$$R = atrt' \left(\frac{1}{1 - r^2} \right)$$

But from Stokes' treatment (Eq. 11y), $t' = 1 - r^2$, so we obtain finally

$$R = ar \quad (14e)$$

This is just equal to the amplitude of the first reflected ray, so we conclude that under the conditions of Eq. 14b there will be complete destructive interference.

If the image of an extended source reflected in a thin plane-parallel film be examined, it will be found to be crossed by a system of distinct interference fringes, provided the source emits monochromatic light and pro-

vided the film is sufficiently thin. Each bright fringe corresponds to a particular path difference giving an integral value of m in Eq. 14c. For any fringe, the value of ϕ is fixed, so the fringe will have the form of the arc of a circle whose center is at the foot of the perpendicular drawn from the eye to the plane of the film. The necessity of using an extended source will become clear upon consideration of Fig. 14A. If a very distant point source S is used, the parallel rays will necessarily reach the eye at only one angle (that required by the law of reflection), and will be focused to a point P . If the lens L is the lens of the eye, P will be on the retina, and only one point will be seen, either bright or dark, according to the phase difference at this particular angle. It is true that, if the source is not very far away, its image on the retina will be slightly blurred, because the eye is focused for parallel rays. The area illuminated is small, however, and in order to see an extended system of fringes, we must obviously have many points S , spread out in a broad source so that the light reaches the eye from various directions.

These fringes are seen by the eye only if the film is very thin, unless the light is reflected practically normal to the film. At other angles, since the pupil of the eye has a small aperture, increasing the thickness of the film will cause the reflected rays to get so far apart that only one enters the eye at a time. Obviously no interference can occur under these conditions. Using a telescope of large aperture, the lens may include enough rays for the fringes to be visible with thick plates, but unless viewed nearly normal to the plate, they will be so finely spaced as to be invisible. The fringes seen with thick plates near normal incidence are called *Haidinger* fringes* and are used in the Fabry-Perot interferometer, to be described later.

14.2. Interference in the Transmitted Light. The rays emerging from the lower side of the film, shown in Figs. 14A and 14C, may also be brought together with a lens and caused to interfere. Here, however, there are no phase changes at reflection for any of the rays, and the relations are such that Eq. 14b now becomes the condition for maxima and Eq. 14c the condition for minima. For maxima the rays u , v , w , . . . of Fig. 14A are in phase, while for minima v , x , . . . are out of phase with u , w , . . . For a film of a substance of low reflecting power like glass or water, where r has a small value, u is very much stronger than the other rays, and the minima are not by any means black.

To derive an equation for the intensity distribution in either the reflected or transmitted fringe system, we must evaluate the sum of an

* W. K. Haidinger (1795–1871). Austrian mineralogist and physicist, for 17 years director of the Imperial Geological Institute in Vienna.

infinite number of vibrations of diminishing amplitude, paying attention to their phase differences. This is not a difficult mathematical problem, but to carry it through would give us little insight into the physical reason for the fringe contours. Hence we shall merely state the result here and make a graphical investigation of the problem later (Sec. 14.8). The intensity no longer varies according to the square of the cosine as with two interfering beams but is given by a more complex expression. For the transmitted fringes*

$$I_T = \frac{(1 - r^2)^2}{1 - 2r^2 \cos \delta + r^4} = \frac{1}{1 + \frac{4r^2 \sin^2 (\delta/2)}{(1 - r^2)^2}} \quad (14f)$$

where r is the fraction of the amplitude reflected at a single surface, so that r^2 is the fraction of the intensity reflected or the *reflecting power*.

As before, δ is the phase difference between successive rays, which from Eq. 14a becomes

$$\delta = \frac{2\pi}{\lambda} \Delta = \frac{4\pi}{\lambda} nd \cos \phi' \quad (14g)$$

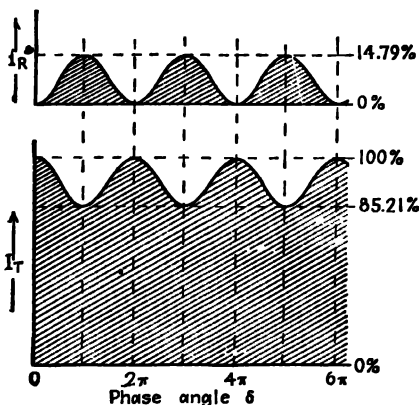


FIG. 14D. Intensity contours for the reflected and transmitted fringes seen by multiple reflection in thin films. Reflecting power 4 per cent.

The upper part of Fig. 14D shows the corresponding intensity contour for the fringes in reflected light. The equation for these need not be written explicitly, since the reflected fringes must be exactly complementary to the transmitted ones.† The amount of light absorbed in transmission through the plate is generally small and can be neglected,

* The derivation of this equation may be found in any standard textbook on optics, for instance in A. Schuster and J. W. Nicholson, "The Theory of Optics," 3d ed., pp. 69-70, Edward Arnold & Co., London, 1924.

† See, however, the qualification to this statement made in Sec 14.8, for the case where absorption is appreciable.

as was done in deriving Eq. 14*f*. The law of conservation of energy then requires that

$$I_R + I_T = 1 \quad (14h)$$

It will be seen in the figure that, while the transmitted fringes are most intense, the reflected ones show much greater contrast between maxima and minima. For either set the interfering rays reaching the eye are parallel (Fig. 14*A*), and the eye or telescope must be focused on infinity as in the case of the circular fringes in the Michelson interferometer.

14.3. Film of Varying Thickness. If the film is not plane-parallel, so that the surfaces make an appreciable angle with each other as in Fig. 14*E*, the interfering rays do not enter the eye parallel to each other, but appear to diverge from a point near the film. The resulting fringes resemble the localized fringes in the Michelson interferometer, and appear to be formed in the film itself. If the two surfaces are plane, so that the film is wedge-shaped, the fringes will be practically straight and parallel to the thin edge of the wedge. In this case the path difference for a given pair of rays is practically that given by Eq. 14*a*. Provided that observations are made almost normal to the film, the factor $\cos \phi'$ may be considered equal to 1, and the condition for bright fringes becomes

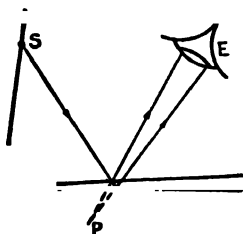


FIG. 14*E*. Illustrating the apparent location of interference fringes as seen by multiple reflection in a wedge-shaped film.

$$2nd = (m + \frac{1}{2})\lambda \quad (14i)$$

In going from one fringe to the next m increases by 1, and this requires that the optical thickness of the film, nd , should change by $\lambda/2$.

Thin film fringes are easily shown in the laboratory or lecture room by using two pieces of ordinary plate glass. If they are laid together with a thin strip of paper along one edge, we obtain a wedge-shaped film of air between the plates. When a sodium flame is viewed as in Fig. 14*E*, yellow fringes are clearly seen. If a carbon arc and filter are used, the fringes may be projected on a screen with a lens.

These fringes from a film of variable thickness have an important practical application in the testing of optical surfaces for planeness. If an air film is formed between two surfaces, one of which is perfectly plane and the other not, the fringes will be irregular in shape. Any

fringe is characterized by a particular value of m in Eq. 14i and hence will follow those parts of the film where d is constant. That is, the fringes form the equivalent of *contour lines* for the uneven surface. The interval between contours is $\lambda/2$, since for air $n = 1$, and going from one fringe to the next corresponds to increasing d by this amount. The standard method of producing optically plane surfaces uses repeated observation of the fringes formed between the working surface and an "optical flat," the polishing being continued until the fringes are straight.

14.4. Newton's* Rings. The celebrated phenomenon known by this name is a special case of interference in a film of variable thickness. We shall consider it in some detail, both because of its historical importance and because it is frequently used in laboratory experiments to measure

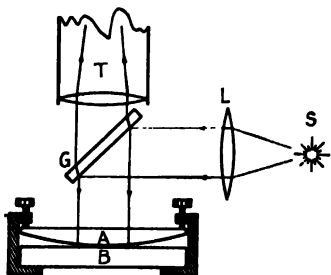


FIG. 14F. Experimental arrangement used in viewing and measuring Newton's rings.

the wavelength of light. An air film is formed by laying the convex side of a plano-convex lens of long focus on a plane glass surface (Fig. 14F). In order to view the fringes at right angles to the film, monochromatic light from the source S is partially reflected from the glass plate G , so that after reflection from the film it enters the observer's eye or the low-power microscope T . Since the air film is symmetrical about the point of contact, the fringes, which follow lines of equal

thickness, will be concentric rings with their center at this point. A photograph of the rings formed with monochromatic light is reproduced in Fig. 14G(a). If white light is used, only a few rings are observed and these will all be highly colored except the central spot, which is black. That the diameter of the rings depends on wavelength is obvious, and the superposition of the rings of all the different wavelengths in white light will clearly give the observed effect.

We shall now derive an equation giving the relation between the radii r of the rings, the wavelength λ , and the radius of curvature R of the convex surface (Fig. 14H). Let d be the thickness of the air film at a

* Isaac Newton (1642–1727). Besides laying foundations of the science of mechanics, Newton devoted considerable time to the study of light and embodied the results in his famous "Opticks." It seems strange that one of the most striking demonstrations of the interference of light, Newton's rings, should be credited to the chief proponent of the corpuscular theory of light. Newton's advocacy of the corpuscular theory was not so uncompromising as it is generally represented. This is evident to anyone consulting his original writings. The original discovery of Newton's rings is now attributed to Robert Hooke.

distance r from the point of contact C . Then d is the so-called *sagitta* of the arc MN , and by Sec. 4.8, is given by

$$d = \frac{r^2}{2R - d} \quad (14j)$$

Since in practice the radius of curvature R is several meters and d a small fraction of a millimeter, we are justified in dropping d from the denominator, so that

$$d = \frac{r^2}{2R} \quad (14k)$$

We shall now observe a dark fringe at P if, according to Eq. 14b,

$$2nd \cos \phi' = m\lambda$$

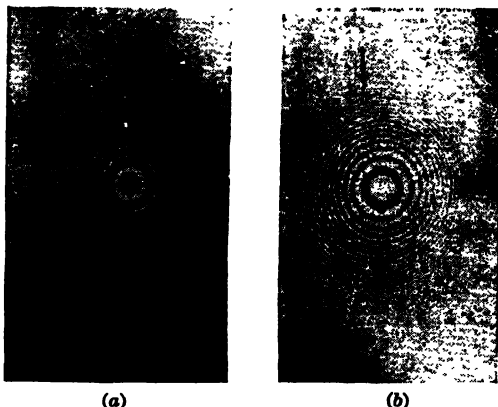


FIG. 14G. Newton's rings (a) by reflection; (b) by transmission.

The angle ϕ' is measured to the normal in the film, and is so nearly zero that we have, very closely,

$$2nd = m\lambda$$

Substituting the resulting value of d in Eq. 14k, we find

$$\frac{m\lambda}{2n} = \frac{r^2}{2R}$$

from which

$$r^2 = \frac{Rm\lambda}{n} \quad \text{DARK RINGS} \quad (14l)$$

For the bright fringes, Eq. 14c gives

$$r^2 = \frac{R(m + \frac{1}{2})\lambda}{n} \quad \text{BRIGHT RINGS} \quad (14m)$$

In an experiment the radii r of the rings may be measured accurately if T in Fig. 14*F* is a traveling microscope. R can be found with a spherometer, and n is known (approximately 1 for air), so we may calculate λ . However, a considerable error is usually made by assuming that the surfaces are tangent at C , as shown in Fig. 14*H*. Even if all dust is eliminated so that the surfaces actually touch, there will be distortion by the pressure of contact. Hence it is more accurate to measure two radii well removed from the center. If r_m is the radius of the dark ring m , and r_{m+s} that of the dark ring $m + s$, we find from Eq. 14*I*, by subtracting the two equations,

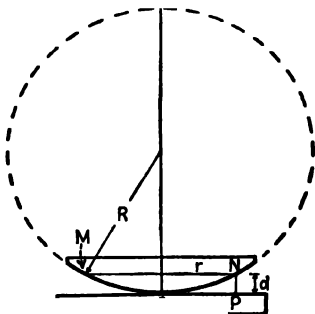


FIG. 14*H*. Geometry of the Newton's-rings experiment.

$$r_{m+s}^2 - r_m^2 = \frac{R(m+s)\lambda}{n} - \frac{Rm\lambda}{n} = \frac{R\lambda s}{n}$$

From this,

$$\lambda = \frac{n(r_{m+s}^2 - r_m^2)}{Rs} \quad (14n)$$

That the central spot of the Newton ring system is black constitutes an experimental proof of the statement in Sec. 14.1 that a relative phase change of π occurs between the rays reflected from glass-to-air and air-to-glass surfaces. If there were no such phase change, the rays reflected from the two surfaces in contact should be in the same phase, and produce a bright spot at the center. In an interesting modification of the experiment, due to Thomas Young, the lower plate has a higher index of refraction than the lens, and the film between is filled with an oil of intermediate index. Then both reflections are at "rare-to-dense" surfaces, no relative phase change occurs, and the central fringe of the reflected system is bright. The experiment does not tell us at which surface the phase change in the ordinary arrangement occurs, but it is now definitely known (Sec. 28.1) that it occurs at the lower (air-to-glass) surface.

A ring system is also observed in the light transmitted by the Newton ring plates. These rings are exactly complementary to the reflected ring system, so that the center spot is now bright. The contrast between bright and dark rings is small, for reasons already discussed in Sec. 14.2. A reproduction of the transmitted pattern is shown in Fig. 14*G*(b).

14.5. Nonreflecting Films. A simple and very important application of the principles of interference in thin films has been the production of the so-called "nonreflecting glass." If a film of a transparent substance

of refractive index n' be deposited on glass of a larger index n , to a thickness of one-quarter of the wavelength of light in the film, so that

$$d = \frac{\lambda}{4n'}$$

the light reflected at normal incidence is almost completely suppressed by interference. This corresponds to the condition $m = 0$ in Eq. 14c, which here becomes a condition for *minima* because the reflections at both surfaces are "rare-to-dense." The waves reflected from the lower surface have an extra path of one-half wavelength over those from the upper surface, and the two, combined with the weaker waves from multiple reflections, therefore interfere destructively. For the destruction to be complete, however, it is necessary that the fraction of the amplitude reflected at each of the two surfaces be exactly the same, since this specification is made in proving the relation of Eq. 14e. It will be true for a film in contact with a medium of higher index only if the index of the film obeys the relation

$$n' = \sqrt{n}$$

This can be proved from Eq. 28b of the chapter on reflection by substituting n' for the refractive index of the upper surface and n/n' for that of the lower. Similar considerations will show that such a film will give zero reflection from the glass side as well as from the air side. Of course no light is destroyed by a nonreflecting film; there is merely a redistribution such that a decrease of reflection carries with it a corresponding increase of transmission.

The practical importance of these films is that by their use one can greatly reduce the loss of light by reflection at the various surfaces of a system of lenses or prisms. Stray light reaching the image as a result of these reflections is also largely eliminated, with a resulting increase in contrast. Almost all optical parts of high quality are now "coated" to reduce reflection. The coatings were first made by depositing several monomolecular layers of an organic substance on glass plates. More durable ones are now made by evaporating calcium or magnesium fluoride on the surface in vacuum, or by chemical treatment with acids which leave a thin layer of silica on the surface of the glass. Properly coated lenses have a purplish hue by reflected light. This is a consequence of the fact that the condition for destructive interference can be fulfilled for only one wavelength, which is usually chosen to be one near the middle of the visible spectrum. The reflection of red and violet light is then somewhat larger. Furthermore, coating materials of sufficient

durability have too high a refractive index to fulfill the condition stated above. Considerable improvement in these respects can be achieved by using two or more superimposed layers, and such films are capable of reducing the total reflected light to one-tenth of its value for the uncoated glass. This refers, of course, to light incident perpendicularly on the surface. At other angles, the path difference will change because of the factor $\cos \phi'$ in Eq. 14c. Since, however, the cosine does not change rapidly in the neighborhood of 0° , the reflection remains low over a fairly large range of angles about the normal. The multiple films may also be used, with suitable thicknesses, to accomplish the opposite purpose—namely, to increase the reflecting power. They may then be used, for example, to divide a light beam into two parts of a given intensity ratio. The division can thus be accomplished without the losses of energy by

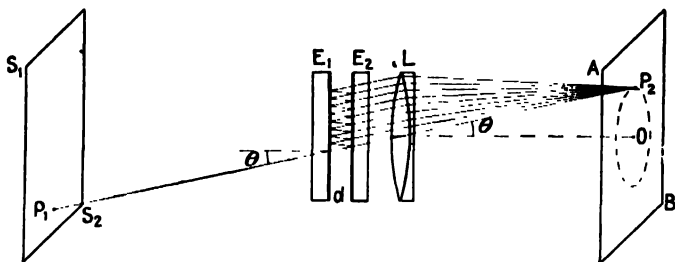


FIG. 14I. Fabry-Perot interferometer E_1E_2 set up to show the formation of circular interference fringes from multiple reflection.

absorption that are inherent in the transmission through, and reflection from, a thin metallic film.

14.6. Fabry-Perot Interferometer. This instrument utilizes the fringes produced in the transmitted light after multiple reflection in the air film between two plane plates thinly silvered on the inner surfaces (Fig. 14I). Since the separation d between the reflecting surfaces is usually fairly large (from 0.1 to 10 cm) and observations are made near the normal direction, the fringes come under the designation of Haidinger fringes mentioned in Sec. 14.1. To observe the fringes, the light from a broad source (S_1S_2) of monochromatic light is allowed to traverse the interferometer plates E_1E_2 . Since any ray incident on the first silvered surface is broken by reflection into a series of *parallel* transmitted rays, it is essential to use a lens L , which may be the lens of the eye, to bring these parallel rays together for interference. In Fig. 14I a ray from the point P_1 on the source is incident at the angle θ , producing a series of parallel rays at the same angle, which are brought together at the point P_2 on the screen AB . It is to be noted that P_2 is *not* an image of P_1 . The

condition for reinforcement of the transmitted rays is given by Eq. 14b with $n = \sqrt{\epsilon}$ for air, and $\phi' = \theta$, so that

$$2d \cos \theta = m\lambda \quad \text{MAXIMA} \quad (14b)$$

This condition will be fulfilled by all points on a circle through P_2 with its center at O , the intersection of the axis of the lens with the screen AB . When the angle θ is decreased, the cosine will increase until another maximum is reached for which n is greater by 1, 2, \dots , so that we have for the maxima a series of concentric rings on the screen with O as their center. Since Eq. 14a is the same as Eq. 13f for the Michelson interferometer, the spacing of the rings is the same as for the circular fringes in that instrument, and they will change in the same way with change in the distance d . In the actual interferometer one plate is fixed, while the other may be moved toward or away from it on a carriage riding on accurately machined ways by a slow-motion screw.

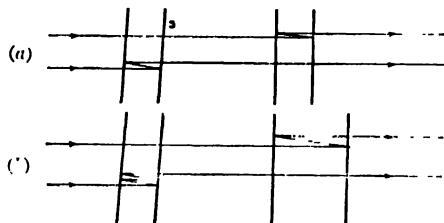


FIG. 14J. Light paths for the formation of Brewster's fringes (a) with two plates of equal thickness; (b) with one plate twice as thick as the other. The inclination of the two plates is exaggerated.

14.7. Brewster's* Fringes. In a single Fabry-Perot interferometer it is not practicable to observe white-light fringes, since the condition of zero path difference occurs only when the two silvered surfaces are brought into direct contact. By the use of two interferometers in series, however, it is possible to obtain interference in white light, and the resulting fringes have had important applications. The two plane-parallel "air plates" are adjusted to exactly the same thickness, or else one to some exact multiple of the other, and the two interferometers are inclined to each other at an angle of 1° or 2° . A ray that bisects the angle between the normals to the two sets of plates can then be split into two, each of which after two or more reflections emerges, having traversed the same path. In Fig. 14J these two paths are drawn as

* Sir David Brewster (1781–1868). Professor of physics at St Andrew's, and later principal of the University of Edinburgh. Educated for the church, he became interested in light through repeating Newton's experiments on diffraction. He made important discoveries in double refraction and in spectrum analysis. Oddly enough, he opposed the wave theory of light in spite of the great advances in this theory that were made during his lifetime.

separate for the sake of clarity, though actually the two interfering beams are derived from the same incident ray, and are superimposed when they leave the system. The reader is referred to Fig. 13Y, where the formation of Brewster's fringes by two thick glass plates in Jamin's interferometer is illustrated. A ray incident at any other angle than that mentioned above will give a path difference between the two emerging ones which increases with the angle, so that a system of straight fringes is produced.

The usefulness of Brewster's fringes lies chiefly in the fact that when they appear, the ratio of the two interferometer spacings is very exactly a whole number. Thus, in the redetermination of the length of the

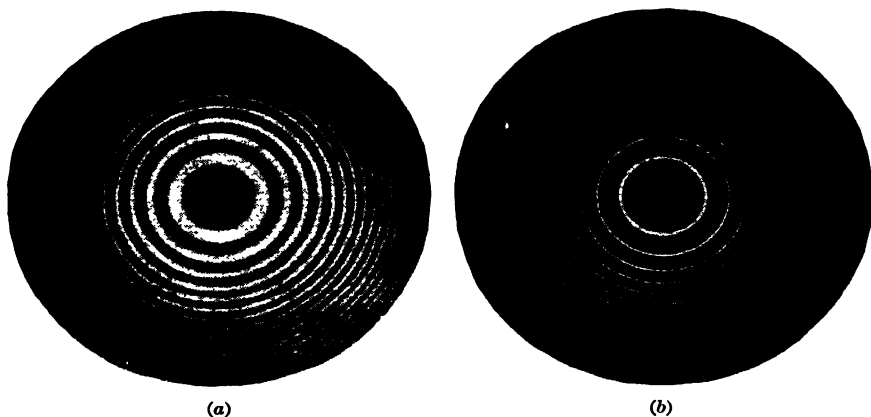


FIG. 14K. Comparison of the types of fringes produced with (a) the Michelson interferometer and (b) the Fabry-Perot interferometer.

standard meter in terms of the wavelength of the red cadmium line, a series of interferometers was made, each having twice the length of the preceding, and these were intercompared using Brewster's fringes. The number of wavelengths in the longest, which was approximately 1 m long, could be found in a few hours by this method. It should finally be emphasized that this type of fringe results from the interference of only *two* beams, and therefore cannot be made very narrow, as can the Haidinger fringes.

14.8. Sharpness of the Haidinger Fringes. If the plates of a Fabry-Perot interferometer are unsilvered, the fringes are broad like those in the Michelson instrument, and their visibility is low. Their intensity contour is that of the lower curve in Fig. 14D. But if the reflecting power of the surfaces is increased by lightly silvering them, a remarkable change in the fringes appears. This is best appreciated by comparing

Fig. 14K(a) and (b). The latter was taken with a Fabry-Perot interferometer whose plates had a reflecting power $r^2 = 0.80$. The intensity is now seen to be concentrated in sharp rings, with relatively broad regions of darkness between them. The intensity in these fringes is governed by Eq. 14f, and the origin of its sharp decrease on either side of a maximum is best seen by examining this expression. When r^2 is large, approaching unity, the term $(1 - r^2)^2$ will be small, and the second term in the denominator will increase very rapidly as $\delta/2$ changes appreciably from its value $m\pi$ for the maxima. In Fig. 14L the intensity contours are plotted from Eq. 14f for $r^2 = 0.04$, 0.5, and 0.8. The curve labeled 4 per cent is just that shown in Fig. 14D, while the two others illustrate

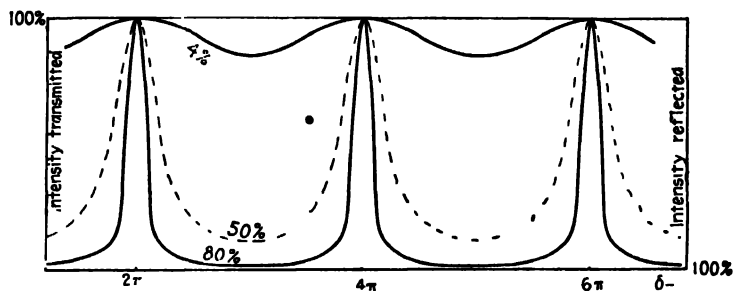


FIG. 14L. Intensity contours for Haidinger fringes, showing how their sharpness depends upon the reflecting power.

the narrowing of the fringes as the reflecting power is increased. The corresponding intensity contours for the Haidinger fringes formed by the *reflected* light, which in the absence of absorption are merely the complement of the transmitted ones, are obtained by inverting the figure or else by inverting the scale as shown at the right.

When the surfaces are thinly silvered, as in the Fabry-Perot interferometer, the relations are not exactly as shown in Fig. 14L because of absorption by the metallic layer. In the first place, the maxima for the transmitted fringes will fall appreciably below the value 100 per cent indicated in the figure. Also, the absorption brings about phase changes in the process of reflection which are neither 0° nor 180° , as will be explained in Sec. 28.7. The result is that the transmitted and reflected fringes are no longer complementary. In fact, if the layer is not too thin, the maxima for the two sets will occur at exactly the same angles. In this case it is no longer legitimate to apply the law of conservation of energy as expressed in Eq. 14h, although the energies integrated over a complete fringe width will be nearly complementary. The reflected light for silvered surfaces of high reflecting power will always give fringes

in which the bright ones are broad and the dark ones narrow, so they are of little practical use.

To understand the narrowness of the transmitted fringes when the reflecting power is high, we may use the graphical method of compounding amplitudes already discussed in Secs. 12.2 and 13.4. Referring back to Fig. 14C we notice that the amplitudes of the transmitted rays are given by att' , $att'r^2$, $att'r^4$, \dots , or in general for the m th ray by $att'r^{2m}$. We thus have to find the resultant of an infinite number of amplitudes which decrease in magnitude more rapidly the smaller the fraction r . In Fig. 14M(a) the magnitudes of the amplitudes of the first 10 transmitted

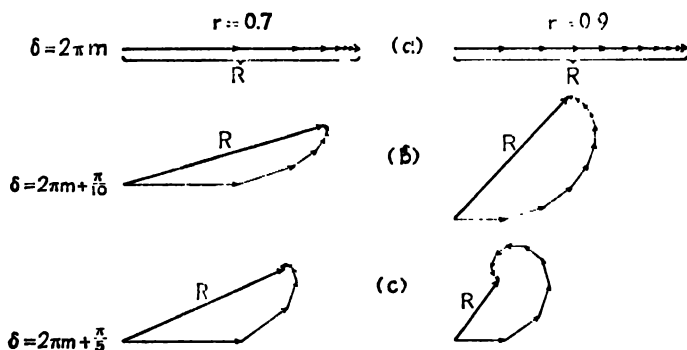


FIG. 14M. Graphical composition of amplitudes for the first 10 rays of the Fabry-Perot interferometer with two different reflecting powers.

rays are drawn to scale for the 50 per cent and 80 per cent cases in Fig. 14L, *i.e.*, essentially for $r = 0.7$ and 0.9 . Starting at any principal maximum, with $\delta = 2\pi m$, these individual amplitudes will all be in phase with each other, so the vectors are all drawn parallel to give a resultant that has been made equal for the two cases. If we now go slightly to one side of the maximum, where the phase difference introduced between successive rays is $\pi/10$, each of the individual vectors must be drawn making an angle of $\pi/10$ with the preceding one, and the resultant found by joining the tail of the first to the head of the last. The result is shown in diagram (b). It will be seen that in the case $r = 0.9$, in which the individual amplitudes are much more nearly equal to each other, the resultant R is already considerably less than in the other case. In diagram (c), where the phase has changed by $\pi/5$, this effect is much more pronounced; the resultant has fallen to a considerably smaller value in the right-hand picture. Although a correct picture would include an infinite number of vectors, the later ones will have vanishing amplitudes, and we would reach a result similar to that found with the first 10.

14.9. Applications of the Fabry-Perot Interferometer and Etalons.

The great advantage of the Fabry-Perot instrument over Michelson's lies in the fact that it gives sharp fringes, so that if different wavelengths are present in the light, they will produce ring systems which are clearly separated. With the diffuse fringes from the Michelson interferometer, this is never possible. The two principal uses to which the Fabry-Perot interferometer has been put are (1) the accurate comparison of the wavelengths of spectral lines, and (2) investigation of the contour of an individual line, including the possibility that it is composed of several component lines very close together, and hence possesses *hyperfine structure*.

The ratio of the wavelengths of two lines, such as the sodium D lines, is sometimes measured in the laboratory with the "sliding interferometer," in which one mirror is movable. Starting with the two mirrors nearly in contact, the ring systems owing to the two wavelengths practically coincide. As d is increased, they gradually separate, and the maximum discordance occurs when the rings of one set are halfway between those of the other set. Confining our attention to the rings at the center ($\cos \theta = 1$), we may write from Eq. 14*b*

$$2d_1 = m_1\lambda = (m_1 + \frac{1}{2})\lambda' \quad (14p)$$

where, of course, $\lambda > \lambda'$. From this,

$$m_1(\lambda - \lambda') = \frac{2d_1}{\lambda} (\lambda - \lambda') = \frac{\lambda'}{2}$$

and

$$\lambda - \lambda' = \frac{\lambda\lambda'}{4d_1} = \frac{\lambda^2}{4d_1}$$

if the difference between λ and λ' is small. On displacing the mirror still farther, the rings will presently coincide and then separate out again. At the next discordance

$$2d_2 = m_2\lambda = (m_2 + 1\frac{1}{2})\lambda' \quad (14q)$$

Subtracting Eq. 14*p* from Eq. 14*q*, we obtain

$$2(d_2 - d_1) = (m_2 - m_1)\lambda = (m_2 - m_1)\lambda' + \lambda'$$

whence, assuming λ approximately equal to λ' , we find

$$\lambda - \lambda' = \frac{\lambda^2}{2(d_2 - d_1)} \quad (14r)$$

We can determine $d_2 - d_1$ either directly from the scale or by counting the number of fringes of the known wavelength λ between discordances.

For the most accurate work, the above method is replaced by one in which the fringe systems of the lines are photographed simultaneously with a fixed separation d of the plates. For this purpose the plates are held rigidly in place by quartz or invar spacers. A pair of Fabry-Perot plates thus mounted is called an etalon (Fig. 14N). The etalon can be used to determine accurately the relative wavelengths of several spectral lines from a single photographic exposure. If it were mounted with a lens as in Fig. 14I, the light containing several wavelengths, the fringe

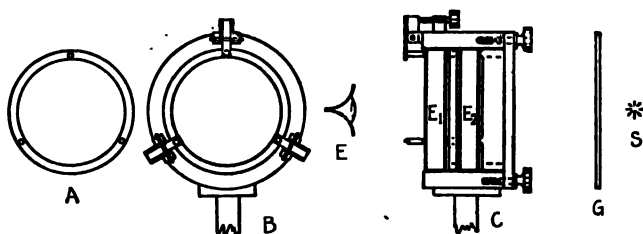


FIG. 14N. Mechanical details of a Fabry-Perot etalon, showing adjustment screws and springs.

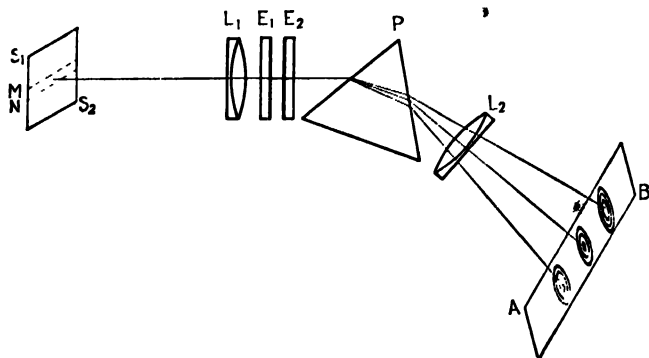


FIG. 14O. Fabry-Perot etalon and prism arrangement for studying small differences in wavelength.

systems of the various wavelengths would be concentric with O and would be confused with each other. However, they can be separated by inserting a prism between the etalon and the lens L . The experimental arrangement is then similar to that shown in Fig. 14O. A photograph of the visible spectrum of mercury taken in this way is shown in the upper part of Fig. 14P. It will be seen that the fringes of the green and yellow lines still overlap. To overcome this, it is merely necessary to use an illuminated slit (MN of Fig. 14O) of the proper width as the source. When the interferometer is in a collimated beam of parallel

light, as it is here, each point on the extended source corresponds to a given point in the ring system. Therefore only vertical sections of the ring system are obtained, as shown in the lower part of Fig. 14P, and these no longer overlap. When the spectrum is very rich in lines, as in Fig. 14Q, the source slit must be made rather narrow. In this photograph only sections of the upper half of the fringe systems appear.

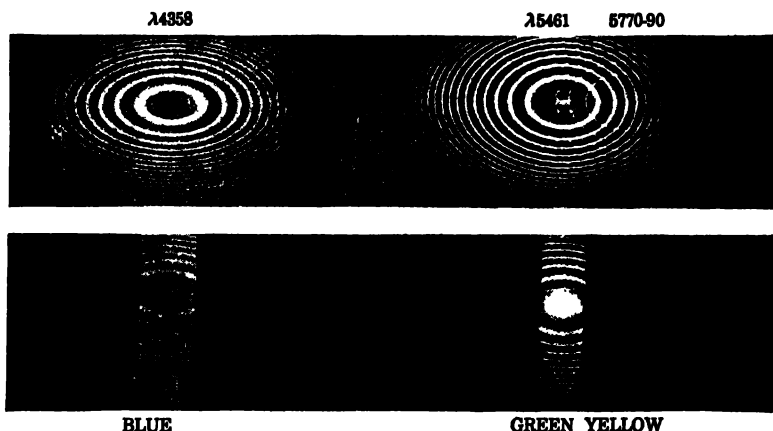


FIG. 14P. Interference rings of the visible mercury spectrum taken with Fabry-Perot etalon as shown in Fig. 14N.

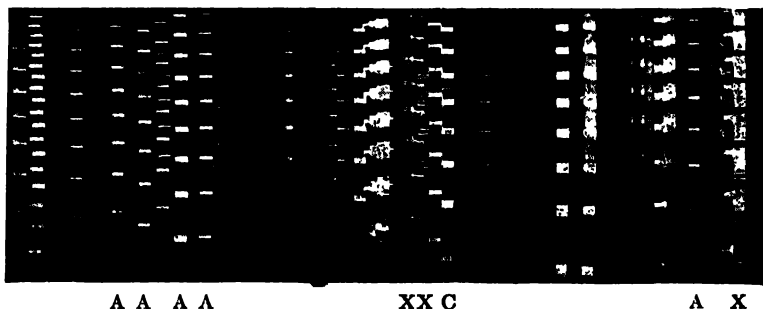


FIG. 14Q. Interference patterns of the lanthanum spectrum taken with a Fabry-Perot etalon. $d = 5$ mm. (After Anderson.)

Measurements of the radii of the rings in a photograph of this type permit very accurate comparison of wavelengths. The determination of the correct values of m in the different systems and of the exact value of d is a rather involved process which we shall not discuss here.* By this method the wavelengths of a hundred lines from the iron arc

* See W. E. Williams, "Applications of Interferometry," 1st ed., pp. 83-88, Methuen & Co., Ltd., London, 1928, for a description of this method.

have been measured relative to the red cadmium line with an accuracy of a few ten-thousandths of an angstrom unit.

The second application of this instrument, the investigation of hyperfine structure, has become of considerable importance in modern research. Occasionally it will be found that a line which appears sharp and single in an ordinary spectroscope will yield ring systems consisting of two or more sets. Examples are found in the lines marked *X* in the lanthanum spectrum (Fig. 14*Q*). Those marked *A* are sharp to a greater or less extent. These multiple ring systems arise from the fact that the line is actually a group of lines of wavelengths very close together, differing by perhaps a few hundredths of an angstrom. If *d* is sufficiently large, these will be separated, so that in each order *m* we obtain effectively a short spectrum very powerfully resolved. Any given fringe of a wavelength λ_1 is formed at such an angle that

$$2d \cos \theta_1 = m\lambda_1 \quad (14s)$$

The next fringe farther out for this same wavelength has

$$2d \cos \theta_2 = (m - 1)\lambda_1 \quad (14t)$$

Suppose now that λ_1 has a component line λ_2 which is very near λ_1 , so that we may write $\lambda_2 = \lambda_1 - \Delta\lambda$. Suppose also that $\Delta\lambda$ is such that this component, in order *m*, falls on the order *m* - 1 of λ_1 . Then

$$2d \cos \theta_2 = m(\lambda_1 - \Delta\lambda) \quad (14u)$$

Equating the right-hand members of Eqs. 14*t* and 14*u*,

$$\lambda_1 = m\Delta\lambda$$

Substituting the value of *m* from Eq. 14*s* and solving for $\Delta\lambda$,

$$\Delta\lambda = \frac{\lambda_1^2}{2d \cos \theta_1} \cong \frac{\lambda_1^2}{2d} \quad (14v)$$

if θ is nearly zero. This is the wavelength interval in a given order when the fringe of the same wavelength in the next higher order is reached. We see that it is constant, independent of *m*. Knowing *d* and λ (approximately), the wavelength difference of component lines lying in this small range may be evaluated.

The frequency of spectrum lines is usually expressed in *wave numbers*, i.e., the number of waves per centimeter path. For any given wavelength, this is just the reciprocal of λ , measured in centimeters,

$$\nu = \frac{1}{\lambda}$$

To find the wave-number difference $\Delta\nu$ corresponding to the $\Delta\lambda$ in Eq. 14v, we may differentiate the above relation to obtain

$$\Delta\nu = -\frac{\Delta\lambda}{\lambda^2}$$

Substitution in Eq. 14v gives

$$\Delta\nu = -\frac{1}{2d} \quad (14w)$$

Hence, if d is expressed in centimeters, $1/2d$ gives the wave-number difference, which is seen to be independent of the order (neglecting the variation of θ) and of wavelength as well.

The smallest wavelength which can be resolved in this way obviously depends upon the width of the maxima of a given λ , relative to the separation of successive orders. As the width decreases, two fringes may be closer together and still be seen as separate. We have seen, as in Fig. 14L, that the width becomes rapidly smaller as the reflecting power of the plates is increased. This figure also shows that the *maxima* of intensity are not changed, so that by using a thicker silver film we obtain sharper fringes without loss of intensity at the maxima. However, this increase in resolving power cannot be continued indefinitely by using thicker and thicker films, for then the intensity of the maxima eventually decreases considerably, owing to *absorption* of light in passing through the silver films. In Eq. 14f, from which this figure was drawn, absorption is neglected (see Sec. 14.8). Finally it should be noted that as $\Delta\lambda$ varies as $1/d$, by Eq. 14v, the smallest wavelength interval resolved is also inversely proportional to d , because the ratio of fringe width to fringe separation is constant. Doubling the separation of the plates therefore doubles the resolving power of this instrument.

The difficult adjustment of the Fabry-Perot interferometer lies in the attainment of accurate parallelism of the silvered surfaces. This operation is usually accomplished by the use of screws and springs, which hold the plates against the spacer rings shown in Fig. 14N. A brass ring *A* with three quartz or invar pins constitutes the spacer. A source of light such as a mercury arc is set up with a sheet of ground glass *G* one side of the etalon, and then viewed from the opposite side as shown in *ECGS*. With the eye focused at infinity, a system of rings will be seen with the reflected image of the pupil of the eye as a center. As the eye is moved up and down or from side to side, the ring system will also move, keeping the eye pupil at the center. If the rings on moving up expand in size, the plates are farther apart at the top than at the bottom.

Tightening the top screw will then depress the corresponding separator pin enough to produce the required change in alignment. When properly adjusted the rings will remain the same size as the eye is moved to any point in the field of view. If the source is bright enough, this adjustment can also be made when the etalon is in the spectrograph, as shown in Fig. 14O.

Sometimes it is convenient to place the etalon in front of the slit of a spectrograph rather than in front of the prism. In such cases the light incident on the etalon need not be parallel. A lens must, however, follow the etalon and this must always be set with the slit at its focal plane. This lens then selects parallel rays from the etalon and focuses interference rings on the slit. Both these methods are used in practice.

14.10. Interference Filter. Our discussion of the Fabry-Perot interferometer has thus far been limited to the dependence of the intensity on plate separation and on angle for a single wavelength, or perhaps for two or more wavelengths close together. If the instrument is placed in a parallel beam of white light, interference will also occur for all the monochromatic components of such light, but this will not manifest itself until the transmitted beam is dispersed by an auxiliary spectroscope. One then observes a series of bright fringes in the spectrum, each formed by a wavelength somewhat different from the next. The maxima will occur, according to Eq. 14o, at wavelengths given by

$$\lambda = \frac{2d \cos \theta}{m} \quad (14x)$$

where m is any whole number. If d is a separation of a few millimeters, there will be very many narrow fringes (more than 12,000 through the visible spectrum when $d = 5$ mm), and high dispersion is necessary in order to separate them. Such fringes have been used for example in the calibration of spectroscopes for the infrared and in accurate measurements of the wavelengths of the absorption lines in the solar spectrum. A very important use for them has recently been found in the case where d is made extremely small, so that only one or two maxima occur within the visible range of wavelengths. With white light incident, only one or two narrow bands of wavelength will then be transmitted, the rest of the light being reflected. The pair of semitransparent metallic films thus can act as a *filter* passing nearly monochromatic light. The curves of transmitted energy against wavelength resemble those of Fig. 14L, since according to Eq. 14g the phase difference δ is inversely proportional to wavelength for a given separation d .

In order that the maxima shall be widely separated, it is necessary that m be a small number. This is attained only by having the reflecting surfaces very close together. If one wishes to have the maximum for $m = 2$ occur at a given wavelength λ , the metal films would have to be a distance of λ apart. The maximum $m = 1$ will then appear at a wavelength of 2λ . Such minute separations can be attained, however, with modern techniques of evaporation in vacuum. A semitransparent metal film is first evaporated on a plate of glass. Next, a thin layer of some dielectric material like quartz is evaporated on top of this, and then the dielectric layer is in turn coated with another similar film of metal. Finally another plate of glass is placed over the films for mechanical protection. The completed filter then has the cross section shown schematically in Fig. 14R, where the thickness of the films is greatly exaggerated relative to that of the glass plates. Since the path difference is now in the dielectric of index n , the wavelengths of maximum transmission for normal incidence are given by

$$\lambda = \frac{2nd}{m} \quad (14y)$$

If there are two maxima in the visible spectrum, one of them can easily be eliminated by using colored glass for the protecting cover plate. Interference filters are now made which transmit a band of wavelengths of width (at half transmission) only 100 Å, with the maximum lying at any desired wavelength. The transmission at the maximum can be as high as 35 per cent. It is very difficult to obtain combinations of colored

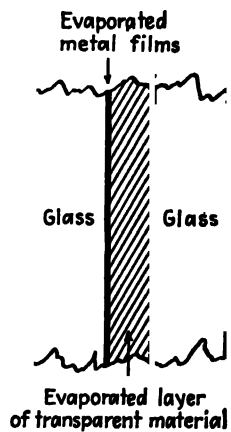


FIG. 14R. Cross section of interference filter.

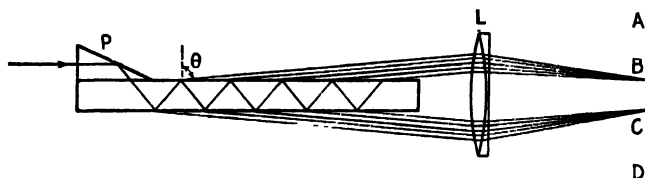


FIG. 14S. Multiple reflection between the surfaces of a Lummer-Gehrcke plate.

glass or gelatin filters which will accomplish this purpose. Furthermore, since the interference filter absorbs a negligible amount of energy, there is no trouble with its overheating.

14.11. Lummer-Gehrcke Plate. Besides the Fabry-Perot interferometer and the Michelson echelon (to be discussed in Sec. 17.17), another

instrument sometimes used for the detailed study of individual spectrum lines consists of an accurately plane-parallel plate of glass or quartz 10 to 20 cm long, 1 or 2 cm wide, and a few millimeters thick. A prism is cemented on one end (P in Fig. 14S) so that the light may enter the plate at the proper angle without excessive loss of intensity by reflection. This angle is such that the angle of incidence on the inner surface is slightly less than the critical angle of total reflection. Thus at each reflection a ray of light leaves the surface at a nearly grazing angle. These rays are parallel, and are brought to a focus by the lens on the screen AD .

The fringes observed with the Lummer-Gehrcke plate are Haidinger fringes observed at an angle θ near 90° , instead of near 0° as in the Fabry-Perot instrument. High reflecting power with consequent sharpness of

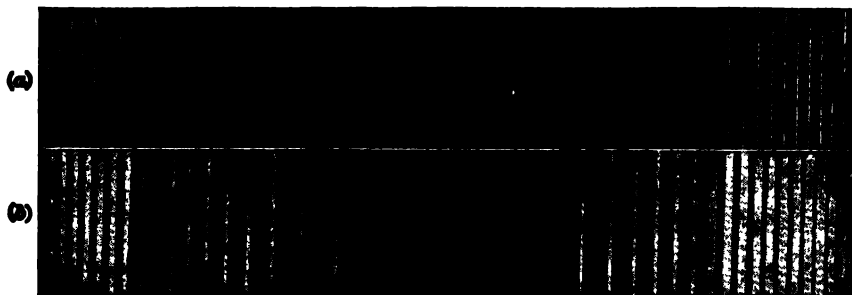


FIG. 14T. Interference fringes from a Lummer-Gehrcke plate, (a) short exposure; (b) long exposure.

the fringes is attained by the fact that near the critical angle the reflection is very strong. On the screen one observes two sets of fringes, one from each side of the plate.* Figure 14T is a photograph of the fringes taken with the green line of mercury. If the line studied is not single but possesses hyperfine structure, it will be resolved in much the same way as described above for the Fabry-Perot interferometer. A Lummer-Gehrcke plate of quartz has an advantage over the Fabry-Perot etalon in that it can be used successfully in the ultraviolet, where the reflecting power of metals is generally low.

Problems

1. Newton's rings are formed by placing the convex surface of a lens in contact with a plane glass surface. If the sixteenth bright ring of green light, $\lambda 5431$, is 12.20 mm in diameter, what is the radius of curvature of the lens? Assume $n = 1.000$ for air.

2. Solve Prob. 1 for the case where the twentieth bright ring has a diameter of 18.45 mm.

* The two fringe systems are not complementary but identical. This is to be expected when the source is effectively *between* the two surfaces (see the discussion in Sec. 14.8).

3. If the radius of curvature of the upper glass surface in a Newton's-ring experiment is 3.5 m, what will be the diameter of the fifth and tenth bright rings for light of the red cadmium line, $\lambda 6438$?

4. Solve Prob. 3 for the green cadmium line, $\lambda 5085$.

5. Two pieces of plane glass are placed together with a piece of paper between the two at one edge. Find the angle in seconds of the wedge-shaped air film between the plates if, on viewing the film with sodium light, $\lambda 5893$, there are 18 fringes per centimeter. Assume that the light is viewed normal to the surface.

6. Solve Prob. 5 if light of the cadmium blue line, $\lambda 4800$, is used.

7. An experiment on Newton's rings is performed with red light and the following measurements are made: $R = 10$ m, radius of m th dark ring = 3.0 mm, radius of $(m + 4)$ th dark ring = 5.0 mm. Find the wavelength of the light used, and the ring number.

8. The reflecting surfaces of a Fabry-Perot etalon are separated by a distance of 4.2 cm. Find the wavelength range $\Delta\lambda$ between adjacent rings when this instrument is used with light of wavelength (a) 4000 Å, (b) 5341 Å, (c) 5893 Å, and (d) 6563 Å.

9. What spacings are required between the two reflecting surfaces of a Fabry-Perot interferometer to make the available wavelength range $\Delta\lambda = 0.0200$ Å for the wavelengths given in Prob. 8?

10. A spectrum line at $\lambda 4750$ is found to be a doublet with a separation of 0.043 Å. What separator in a Fabry-Perot etalon will give a regular set of interference rings with each ring m of one component superimposed on the $(m + 1)$ st ring of the other component?

11. If in taking the photograph in Fig. 14Q the Fabry-Perot separator used had a thickness of 3.2 cm, and the left-hand line marked X had a wavelength of 4800 Å, what would be the difference in wavelength between the first two strong components of the six-line pattern? (NOTE: Use dividers and a scale, or a comparator, to measure the distances between components.)

12. A Lummer-Gehrcke plate 8 mm thick is to be used for studying the red cadmium line, $\lambda 6438$. If the index of refraction is 1.562, find the order of interference occurring nearest to the faces of the plate.

13. Find graphically the intensity due to the first five transmitted rays from a Fabry-Perot interferometer where the phase difference between successive rays is 45° , and express the result relative to that for zero phase difference. Assume a reflecting power of 0.81.

14. Interference fringes are produced in a thin wedge-shaped film of cellophane of index 1.4. If the angle of the film is 20 seconds of arc, and the distance between fringes is 0.25 cm, find the wavelength of the light. Assume perpendicular incidence.

15. The reflecting power of the silvered surfaces of a Fabry-Perot etalon is 64 per cent. Find the minimum intensity, halfway between the maxima of the transmitted fringes.

16. The reflecting power of the silvered surfaces of a Fabry-Perot etalon is 85 per cent. Find the ratio of intensity between the maxima and minima of the transmitted fringes, assuming no absorption.

17. An interference filter is to be designed using a dielectric of index 1.420 as the separator. It is desired that the filter have a maximum transmission at 6000 Å, and a width at half-maximum of 200 Å. The filter is to incorporate a yellow glass which will suppress the next maximum at 5000 Å. Find (a) the necessary thickness of the dielectric, and (b) the required reflecting power of either of the metal films.

18. What are the wavelengths of the maxima in the light transmitted by a piece of

cellophane 0.0013 in. thick, and lightly silvered on both sides? Assume the index 1.40 to be independent of wavelength.

19. A strong line having $\lambda = 5543.02 \text{ \AA}$ and its weak satellite line form systems of fringes which coincide for a particular setting of a sliding Fabry-Perot interferometer. While the separation of the mirrors is slowly increased, 140,100 fringes of the strong line are counted as they appear at the center, and the two fringe systems then coincide again. What is the wavelength difference between the line and its satellite?

20. Eq. 14r applies to the case where $\lambda - \lambda'$ is small. Derive a similar equation applicable for the general case where $\lambda - \lambda'$ need not be small.

21. The surfaces of a prism of refractive index 1.52 are to be made "nonreflecting" by coating them with a thin layer of transparent material of refractive index 1.30. Take the effective wavelength of the light (in vacuum) to be 5500 \AA . (a) Find the necessary thickness of the layer. (b) What is the phase difference between the light reflected from the upper and lower surfaces for violet light, $\lambda 4000$? (c) What is it for red light, $\lambda 7000$?

22. In measuring the coefficient of thermal expansion of a rare material, the sample has two parallel faces about 1 cm apart, and rests upon one of these on a flat surface. A plane glass plate is then placed over it, supported by two copper blocks which are also about 1 cm high but which leave a thin air film between the lower glass surface and the upper surface of the sample. Interference fringes are produced by reflecting sodium light, $\lambda 5890$, from this film, and these fringes are observed in a low-power microscope, which is first set on a dark fringe. On raising the temperature of the whole system by 100°C , 20 dark fringes (not including the first) are counted before the system comes to equilibrium with a dark fringe on the cross hairs. If the coefficient of expansion of copper is 14×10^{-6} per degree centigrade, what is the coefficient of expansion of the sample?

CHAPTER 15

FRAUNHOFER DIFFRACTION BY A SINGLE OPENING

When a beam of light passes through a narrow slit, it spreads out to a certain extent into the region of the geometrical shadow. This effect, already noted and illustrated at the beginning of Chaps. 1 and 13, is one of the simplest examples of *diffraction*, *i.e.*, of the failure of light to travel in straight lines. It can be satisfactorily explained only by assuming a wave character for light, and in this chapter we shall investigate quantitatively the *diffraction pattern*, or distribution of intensity of the light behind the aperture, using the principles of wave motion already discussed.

15.1. Fresnel and Fraunhofer Diffraction. Diffraction phenomena are conveniently divided into two general classes, (1) those in which the source of light and the screen on which the pattern is observed are effectively at infinite distances from the aperture causing the diffraction, and (2) those in which either the source or the screen, or both, are at finite

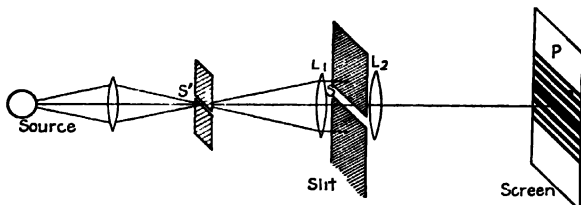


FIG. 15A. Experimental arrangement for obtaining the diffraction pattern of a single slit. Fraunhofer diffraction.

distances from the aperture. The phenomena coming under class (1) are called, for historical reasons, *Fraunhofer diffraction*, and those coming under class (2) *Fresnel diffraction*. Fraunhofer diffraction is much simpler to treat theoretically. It is easily observed in practice by rendering the light from a source parallel with a lens, and focusing it on a screen with another lens placed behind the aperture, an arrangement which effectively removes the source and screen to infinity. In the observation of Fresnel diffraction, on the other hand, no lenses are necessary, but here the wave fronts are divergent instead of plane, and the theoretical treatment is consequently more complex. Only Fraunhofer diffraction will be considered in this chapter.

15.2. Diffraction by a Single Slit. A slit is a rectangular aperture of length large compared to its breadth. Consider a slit S to be set up as in Fig. 15A, with its long dimension perpendicular to the plane of the page, and to be illuminated by parallel monochromatic light from the narrow slit S' , at the principal focus of the lens L_1 . The light focused by another lens L_2 on a screen or photographic plate P at its principal focus will form a diffraction pattern, as indicated schematically. Figure 15B(b) and (c) shows two actual photographs, taken with different exposure times, of such a pattern, using violet light of wavelength 4358 Å. The distance $S'L_1$ was 25 cm, and L_2P was 100 cm. The width of the

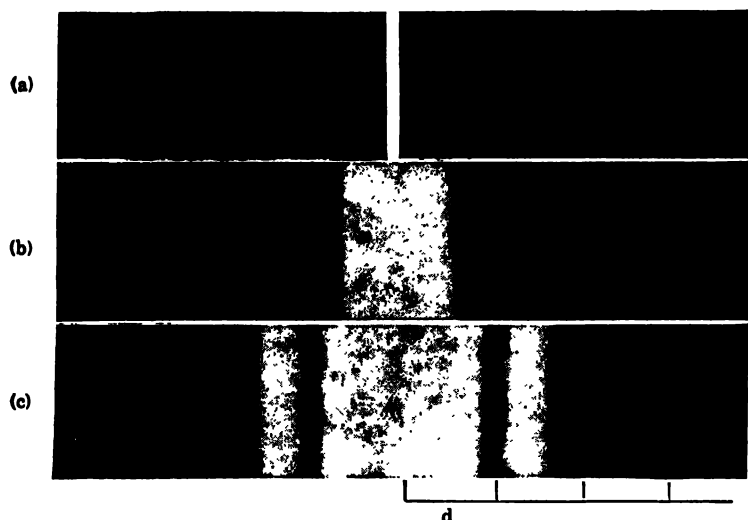


FIG. 15B. Photographs of the single-slit diffraction pattern.

slit S was 0.090 mm, and of S' , 0.10 mm. If S' was widened to more than about 0.3 mm, the details of the pattern began to be lost. On the original plate, half width d of the central maximum was 4.84 mm. It is important to notice that the width of the central maximum is *twice* as great as that of the fainter side maxima. That this effect comes under the heading of diffraction as previously defined is clear when we note that the strip drawn in Fig. 15B(a) is the width of the geometrical image of the slit S' , or practically that which would be obtained by removing the second slit and using the whole aperture of the lens. This pattern can easily be observed by ruling a single transparent line on a photographic plate and using it in front of the eye as explained in Sec. 13.2.

The origin of the single-slit diffraction pattern appears when we investigate the interference of the secondary wavelets which, by Huygens' principle, can be thought of as sent out by every point on a wave front at the instant it occupies the plane of the slit. Figure 15C represents a section of a slit of width a , illuminated with parallel light from the left. Let ds be an element of width of the wave front in the plane of the slit, at a distance s from the center O which we shall call the origin. Each secondary wavelet can be regarded as a spherical wave spreading out to the right, and the parts of each wave traveling normal to the plane of the slit will reach the point P_0 . The parts traveling at the angle θ will reach P_n .

Considering first the wavelet emitted by the element ds situated at the origin, its amplitude will be directly proportional to the length ds ,

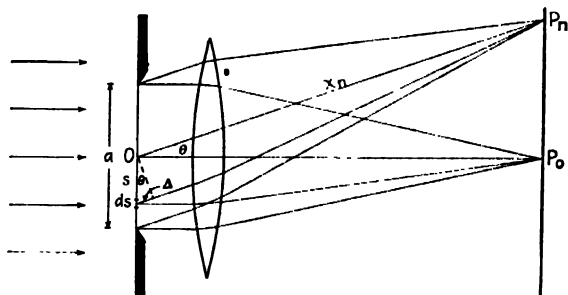


FIG. 15C. Geometrical construction for investigating the intensity in the single-slit diffraction pattern.

and inversely proportional to the distance x . If dy represents the displacement in the wave front, we may represent it by the equation of a spherical wave (Eq. 11*u*) in the form

$$dy = \frac{r ds}{x} \sin 2\pi \left(\frac{t}{T} - \frac{x}{\lambda} \right) \quad (15a)$$

Here $r ds$ is the amplitude of the wave at unit distance from the origin. The part of this wavelet reaching P_n traverses an optical path x_n , and produces a vibration at P_n given by

$$dy_0 = \frac{r ds}{x_n} \sin 2\pi \left(\frac{t}{T} - \frac{x_n}{\lambda} \right) \quad (15b)$$

The phases of the contributions from the other wavelets will be different, since each travels a different distance to P_n . For the element ds at a dis-

tance s below the origin, the wave travels an additional distance $\Delta = s \sin \theta$, and its contribution at P_n may be expressed as

$$\begin{aligned} dy_n &= \frac{r}{x} ds \sin 2\pi \left(\frac{t}{T} - \frac{x_n + \Delta}{\lambda} \right) \\ &= \frac{r}{x} ds \sin 2\pi \left(\frac{t}{T} - \frac{x_n}{\lambda} - \frac{s \sin \theta}{\lambda} \right) \end{aligned} \quad (15c)$$

We now wish to sum up the contributions of all elements ds from one edge of the slit to the other to get the resultant displacement at P_n . This sum is obtained by integrating the expression over the width of the slit, *i.e.*, between the limits $s = -(a/2)$, and $s = +(a/2)$. In doing so, we may drop the amplitude factor r/x , since it is practically the same for all wavelets, and we are interested only in *relative* intensities on the screen.

Adopting the abbreviations $\phi = 2\pi \left(\frac{t}{T} - \frac{x_n}{\lambda} \right)$ and $\psi = 2\pi \frac{s \sin \theta}{\lambda}$, the integral is

$$y = \int_{-a/2}^{+a/2} \sin(\phi - \psi) ds \quad (15d)$$

$$\begin{aligned} &= \int_{-a/2}^{+a/2} (\sin \phi \cos \psi - \cos \phi \sin \psi) ds \\ &= \left[s \frac{\sin \psi}{\psi} \sin \phi + \frac{s \cos \psi}{\psi} \cos \phi \right]_{-a/2}^{+a/2} \end{aligned} \quad (15e)$$

The values of ψ for $s = +a/2$ and $-a/2$ are $(\pi a \sin \theta)/\lambda$ and $(-\pi a \sin \theta)/\lambda$, respectively. Making the required substitutions, one finds

$$y = a \frac{\sin [(\pi a \sin \theta)/\lambda]}{(\pi a \sin \theta)/\lambda} \sin 2\pi \left(\frac{t}{T} - \frac{x_n}{\lambda} \right) \quad (15f)$$

as the equation for the resultant vibration. This represents a new simple periodic motion, of phase differing by $-2\pi x_n/\lambda$ from that at the slit, and of amplitude $R = (a \sin \beta)/\beta$, where $\beta = (\pi/\lambda)a \sin \theta$. Since this expression for R differs by some arbitrary constant from the true amplitude because of the neglect of the factor r/x , it is more correct to write

$$R = R_0 \frac{\sin \beta}{\beta}, \quad \left(\beta = \frac{\pi a \sin \theta}{\lambda} \right) \quad (15g)$$

The significance of the constant R_0 will appear below. The quantity β is a convenient variable, and signifies one-half the phase difference in radians between the contributions from opposite edges of the slit. It determines the intensity by the relation

$$\text{Intensity } I = R^2 = R_0^2 \frac{\sin^2 \beta}{\beta^2} \quad (15h)$$

If the light, instead of being incident on the slit perpendicular to its plane, makes an angle i , a little consideration will show that it is merely necessary to replace the above expression for β by the more general expression

$$\beta = \frac{\pi a(\sin i + \sin \theta)}{\lambda} \quad (15i)$$

15.3. Further Investigation of the Single-slit Diffraction Pattern. In Fig. 15D(a) graphs are shown of Eq. 15g for the *amplitude* (dotted curve) and Eq. 15h for the *intensity*, taking the constant R_0 in each case as unity. The intensity curve will be seen to have the form required by the experimental result in Fig. 15B. The maximum intensity of the

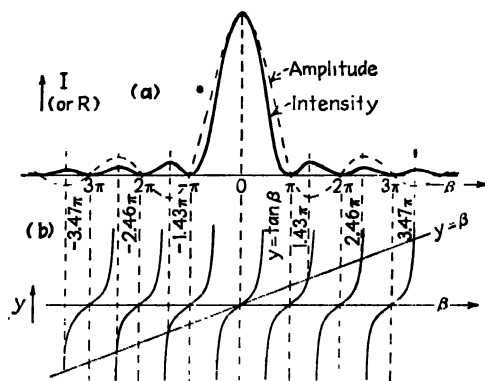


FIG. 15D. Amplitude and intensity contours for Fraunhofer diffraction of a single slit, showing positions of maxima and minima.

strong central band comes at the point P_0 of Fig. 15C, where evidently all the secondary wavelets will arrive in phase because the path difference $\Delta = 0$. For this point $\beta = 0$, and although the quotient $(\sin \beta)/\beta$ becomes indeterminate for $\beta = 0$, it will be remembered that $\sin \beta$ approaches β for small angles, and is equal to it when β vanishes. Hence for $\beta = 0$, $(\sin \beta)/\beta = 1$. We now see the significance of the constant R_0 . Since for $\beta = 0$, $R = R_0$, it represents the amplitude when all the wavelets arrive in phase. R_0^2 is then the value of the maximum intensity, at the center of the pattern. From this *principal maximum* the intensity falls to zero at $\beta = \pm\pi$, then passes through several *secondary maxima*, with equally spaced points of zero intensity at $\beta = \pm\pi, \pm 2\pi, \pm 3\pi, \dots$, or in general $\beta = m\pi$. The secondary maxima do not fall halfway between these points, but are displaced toward the center of the pattern by an amount which decreases with increasing m . The

exact values of β for these maxima can be found by differentiating Eq. 15h with respect to β and equating to zero. This yields the condition

$$\tan \beta = \beta \quad (15j)$$

The derivation of this condition is left as a problem for the student (see Prob. 12 at the end of this chapter). The values of β satisfying this relation are easily found graphically as the intersections of the curve $y = \tan \beta$ and the straight line $y = \beta$. In Fig. 15D(b) these points of intersection lie directly below the corresponding secondary maxima.

The intensities of the secondary maxima may be calculated to a very close approximation by finding the values of $(\sin^2 \beta)/\beta^2$ at the halfway positions, *i.e.*, where $\beta = 3\pi/2, 5\pi/2, 7\pi/2, \dots$. This gives $4/(9\pi^2)$, $4/(25\pi^2)$, $4/(49\pi^2)$, \dots , or $\frac{1}{22.2}, \frac{1}{61.7}, \frac{1}{121}, \dots$, of the intensity of the principal maximum. The greatest error in these values occurs for

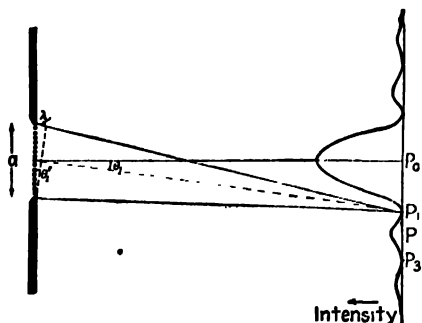


FIG. 15E. Angle of the first minimum of the single-slit diffraction pattern.

the first secondary maximum, which is given as 4.50 per cent of the central intensity, compared with its true value of 4.72 per cent.

A very clear idea of the origin of the single-slit pattern is obtained by the following simple treatment. Consider the light from the slit of Fig. 15E coming to the point P_1 on the screen, this point being just one wavelength farther from the upper edge of the slit than from the lower.

The secondary wavelet from the point in the slit adjacent to the upper edge will travel approximately $\lambda/2$ farther than that from the point at the center, and so these two will produce vibrations with a phase difference of π and will give a resultant displacement of zero at P_1 . Similarly the wavelet from the next point below the upper edge will cancel that from the next point below the center, and we may continue this pairing off to include all points in the wave front, so that the resultant effect at P_1 is zero. At P_2 the path difference is 2λ , and if we divide the slit into four parts, the pairing of points again gives zero resultant, since the parts cancel in pairs. For the point P_3 , on the other hand, the path difference is $3\lambda/2$, and we may divide the slit into thirds, two of which will cancel, leaving one third to account for the intensity at this point. The resultant amplitude at P_2 is, of course, not even approximately

one-third that at P_0 , because the phases of the wavelets from the remaining third are not by any means equal.

The above method, though instructive, is not rigorous if the screen is at a finite distance from the slit. As Fig. 15E is drawn, the shorter broken line is drawn to cut off equal distances on the rays to P_1 . It will be seen from this that the path difference to P_1 between the light coming from the upper edge and that from the center is slightly greater than $\lambda/2$, and that between the center and lower edge slightly less than $\lambda/2$. Hence the resultant intensity will not be zero at P_1 and P_3 , but it will be more nearly so the greater the distance between slit and screen, or the narrower the slit. This corresponds to the transition from Fresnel diffraction to Fraunhofer diffraction. Obviously, with the relative dimensions shown in the figure, the geometrical shadow of the slit would considerably widen the central maximum as drawn. When the screen is at infinity, the relations become simpler, for then the two angles θ_1 and θ'_1 in Fig. 15E become exactly equal (*i.e.*, the two dashed lines are perpendicular to each other), and $\lambda = a \sin \theta_1$ for the first minimum. In practice θ_1 is usually a very small angle, so we may put the sine equal to the angle. Then

$$\theta_1 = \frac{\lambda}{a} \quad (15k)$$

a relation which shows at once how the dimensions of the pattern vary with λ and a . The *linear* width of the pattern on a screen will be proportional to the slit-screen distance, which is the focal length f' of a lens placed close to the slit. The linear distance d between successive minima corresponding to the angular separation $\theta_1 = \lambda/a$ is thus

$$d = \frac{f\lambda}{a} \quad (15l)$$

The width of the pattern increases in proportion to the wavelength, so that for red light it is roughly twice as wide as for violet light, the slit width, etc., being the same. If white light is used, the central maximum is white in the middle, but is reddish on its outer edge, shading into a purple and other impure colors farther out.

The angular width of the pattern for a given wavelength is inversely proportional to the slit width a , so that as a is made larger, the pattern shrinks rapidly to a smaller scale. In photographing Fig. 15B, if the slit S had been 9 mm wide, the whole visible pattern (of five maxima) would be included in a width of 0.24 mm on the original plate instead of 2.4 cm. This fact, that when the width of the aperture is large compared

to a wavelength the diffraction is practically negligible, led the early investigators to conclude that light travels in straight lines and that it could not be a wave motion. Sound waves, whose lengths are measured in feet, will evidently be diffracted through large angles in passing through an aperture of ordinary size, such as an open window.

Finally, it should be emphasized that Eq. 15*h* for the intensity is not exact in two respects. In the first place, a more rigorous treatment* shows that the amplitude should be multiplied by a factor $1 + \cos \theta$, hence the intensity by $(1 + \cos \theta)^2$. This is the so-called *obliquity factor*, which is usually neglected in dealing with these problems, since θ is a very small angle in most cases. In the second place, the equation does not hold when the slit width a becomes less than one wavelength. With $a = \lambda$, $\sin \theta_1 = 1$, so that the first minimum occurs at 90° , and no secondary maxima can occur. The light which would go into these cannot be destroyed, however, and the intensity formula should be modified to account for it elsewhere.

15.4. Graphical Treatment of Amplitudes. The Vibration Curve. The addition of the amplitude contributions from all the secondary wavelets originating in the slit may be carried out by a graphical method based on the vector addition of amplitudes discussed in Sec. 12.2. It will be worth while to consider this method in some detail, because it may be applied to advantage in other more complicated cases to be treated in later chapters, and because it gives a very clear physical picture of the origin of the diffraction pattern. Let us divide the width of the slit into a fairly large number of equal parts, say 10. The amplitude r contributed at a point on the screen by any one of these parts will be the same, since they are of equal width. The phases of these contributions will differ, however, for any point except that lying on the axis, *i.e.*, on the normal to the slit at its center (P_0 , Fig. 15*C*). For a point off the axis, each of the 10 segments will contribute vibrations differing in phase, because the segments are at different average distances from the point. Furthermore the difference in phase δ between the contributions from adjacent segments will be constant, since each element is on the average the same amount farther away (or nearer) than its neighbor.

Now, using the fact that the resultant amplitude and phase may be found by the vector addition of the individual amplitudes making angles with each other equal to the phase difference, a vector diagram like that

shown in Fig. 15F(b) may be drawn. Each of the 10 equal amplitudes r is inclined at an angle δ with the preceding one, and their vector sum R is the resultant amplitude required (see also Fig. 12B). Now suppose that instead of dividing the slit into 10 elements, we had divided it into many thousand or, in the limit, an infinite number of equal elements. The vectors r would become shorter, but at the same time δ would decrease in the same proportion, so that in the limit our vector diagram would approach the arc of a circle, shown as in (b'). The resultant amplitude R is still the same and equal to the length of the chord of this arc. Such a continuous curve, representing the addition of infinitesimal amplitudes, we shall refer to as a *vibration curve*.

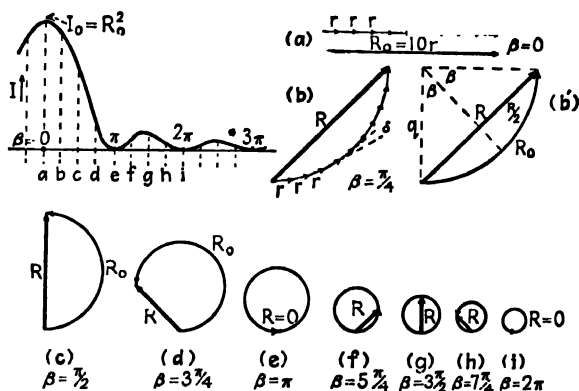


FIG. 15F. Graphical treatment of amplitudes in single-slit diffraction.

To show that this method is in agreement with our previous result, we note that the length of the arc is just the amplitude R_0 obtained when all of the component vibrations are in phase, as in (a) of the figure. Introducing a phase difference between the components does not alter their individual amplitudes or the algebraic sum of these. Hence the ratio of the resultant amplitude R at any point in the screen to R_0 , that on the axis, is the ratio of the chord to the arc of the circle. In terms of our chosen variable β (Sec. 15.2), which represents $(\pi a \sin \theta)/\lambda$, or half the phase difference from opposite edges of the slit, the angle subtended by the arc is just 2β , because the first and last vectors r will have a phase difference of 2β . In Fig. 15F(b'), the radius of the arc is called q , and a perpendicular has been dropped from the center on the chord R . From the geometry of the figure, we have

$$\sin \beta = R/2q \qquad R = 2q \sin \beta$$

and hence

$$\frac{R}{R_0} = \frac{\text{chord}}{\text{arc}} = \frac{2q \sin \beta}{q \times 2\beta} = \frac{\sin \beta}{\beta}$$

in agreement with Eq. 15*g*.

As we go out from the center of the diffraction pattern, the length of the arc remains constant and equal to R_0 , but its curvature increases owing to the larger phase difference δ introduced between the infinitesimal component vectors r . The vibration curve thus winds up on itself as β is increased. The successive diagrams (a) to (i) in Fig. 15*F* are drawn for the indicated values of β at intervals of $\pi/4$, and the corresponding points are similarly lettered on the intensity diagram. A study of these figures will bring out clearly the cause of the variations in intensity occurring in the single-slit pattern.

15.5. Rectangular Aperture. In the preceding sections the intensity function for a slit was derived by summing the effects of the spherical wavelets originating from a linear section of the wave front by a plane perpendicular to the length of the slit, *i.e.*, by the plane of the page in Fig. 15*C*. Nothing was said about the contributions from parts of the wave front out of this plane. A more thorough mathematical investigation, involving a double integration over both dimensions of the wave front,* shows, however, that the above result is correct when the slit is very long compared to its width. The complete treatment gives, for a slit of width a and length l , the following expression for the intensity:

$$I = \text{const} \cdot a^2 l^2 \frac{\sin^2 \beta}{\beta^2} \frac{\sin^2 \gamma}{\gamma^2} \quad (15m)$$

where $\beta = (\pi a \sin \theta)/\lambda$, as before, and $\gamma = (\pi l \sin \Omega)/\lambda$. The angles θ and Ω are measured from the normal to the aperture at its center, in planes through the normal parallel to the sides a and l , respectively. The diffraction pattern given by Eq. 15*m* when a and l are comparable with each other is illustrated in Fig. 15*G*. The dimensions of the aperture are shown by the white rectangle in the lower left-hand part of the figure. The intensity in the pattern is concentrated principally in two directions coinciding with the sides of the aperture, and in each of these directions it corresponds to the simple pattern for a slit width equal to the width of the aperture in that direction. Owing to the inverse proportionality between the slit width and the scale of the pattern, the fringes are more closely spaced in the direction of the longer dimension of the aperture. In addition to these patterns there are other faint

* See R. W. Wood, "Physical Optics," 2d ed., pp. 195-202, The Macmillan Company, New York, 1921.

maxima, as shown in the figure. This diffraction pattern may easily be observed by illuminating a small rectangular aperture with monochromatic light from a source which is effectively a *point*, the disposition of the lenses and the distance of the source and screen being similar to those described for observation of the slit pattern (Sec. 15.2). The cross formed by the brightest spots in the photograph is the one always observed when a bright street light is seen through a wet umbrella.

Now for a slit having l very large, the factor $(\sin^2 \gamma)/\gamma^2$ in Eq. 15*m* is zero for all values of Ω except extremely small ones. This means that the diffraction pattern will be limited to a line on the screen perpendicu-



FIG. 15*G*. Diffraction pattern from a rectangular opening. (After A. Köhler.)

lar to the slit and will resemble a section of the central horizontal line of bright spots in Fig. 15*G*. We do not ordinarily observe such a line pattern in diffraction by a slit, because its observation requires the use of a *point source*. In Fig. 15.1 the primary source was a slit S' , with its long dimension perpendicular to the page. In this case, each point of the source slit forms a line pattern, but these fall adjacent to each other on the screen, adding up to give a pattern like Fig. 15*B*. If we were to use a slit source with the rectangular aperture of Fig. 15*G*, the slit being parallel to the side l , the result would be the summation of a number of such patterns, one above the other, and would be identical with Fig. 15*B*.

These considerations can easily be extended to cover the effect of widening the primary slit. With a slit of finite width, each line element parallel to the length of the slit forms a pattern like Fig. 15*B*. The resultant pattern is equivalent to a set of such patterns displaced laterally with respect to each other. If the slit is too wide, the single-slit pattern will therefore be lost. No great change will occur until the patterns

from the two edges of the slit are displaced about one-fourth of the distance from the central maximum to the first minimum. This condition will hold when the width of the primary slit subtends an angle of $\frac{1}{4}(\lambda/a)$ at the first lens, as can be seen by reference to Fig. 15II below.

15.6. Resolving Power with a Rectangular Aperture. By the resolving power of any optical system we mean its ability to produce separate images of objects very close together. Using the laws of geometrical optics, a telescope or a microscope is designed to give an image of a point source which is as small as possible. However, in the final analysis, it is the diffraction pattern that sets a theoretical upper limit to the resolving power. We have seen that whenever parallel light passes through any aperture, it cannot be focused to a point image, but instead gives a diffraction pattern in which the central maximum has a certain

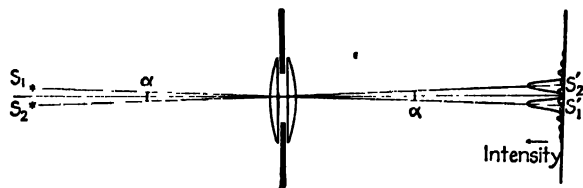


FIG. 15II. Diffraction images of two slit sources by a rectangular aperture.

finite width, inversely proportional to the width of the aperture. The images of two objects will evidently not be resolved if their separation is much less than the width of the central diffraction maximum. The aperture here involved is usually that of the objective lens of the telescope or microscope and is therefore circular. Diffraction by a circular aperture will be considered below in Sec. 15.8, and here we shall treat the somewhat simpler case of a rectangular aperture.

Figure 15II shows two plano-convex lenses (equivalent to a single double-convex lens) limited by a rectangular aperture of vertical dimension a . Two narrow slit sources S_1 and S_2 perpendicular to the plane of the figure form real images S'_1 and S'_2 on a screen. Each image consists of a single-slit diffraction pattern for which the intensity distribution is plotted in a vertical direction. The angular separation α of the central maxima is equal to the angular separation of the sources, and with the value shown in the figure is adequate to give separate images. The condition illustrated is that in which each principal maximum falls exactly on the second minimum of the adjacent pattern. This is the smallest possible value of α which will give zero intensity between the two strong maxima in the resultant pattern. The angular separation from the center to the second minimum in either pattern then corresponds

to $\beta = 2\pi$ (see Fig. 15D), or $\sin \theta \cong \theta = 2\lambda/a = 2\theta_1$.^{*} As α is made smaller than this, and the two images move closer together, the intensity between the maxima will rise, until finally no minimum remains at the center. Figure 15I illustrates this by showing the resultant curve (heavy line) for four different values of α . In each case the resultant pattern is obtained by merely adding the intensities due to the separate patterns (dotted and light curves).

Inspection of this figure shows that it would be impossible to resolve the two images if the maxima were much closer than $\alpha = \theta_1$, corresponding to $\beta = \pi$. At this separation the maximum of one pattern falls exactly on the first minimum of the other, so that the intensities of the

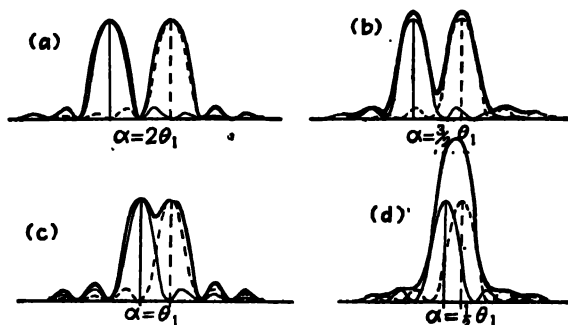


FIG. 15I. Diffraction images of two slit sources: (a) and (b) well resolved; (c) just resolved; (d) not resolved.

maxima in the resultant pattern are equal to those of the separate maxima. To find the intensity at the center of the resultant minimum, we note that the curves cross at $\beta = \pi/2$ for either pattern and

$$\frac{\sin^2 \beta}{\beta^2} = \frac{4}{\pi^2} = 0.4053$$

the intensity of either relative to the maximum. The sum of the contributions at this point is therefore 0.8106, which shows that the intensity of the resultant pattern drops almost to four-fifths of its maximum value. This change of intensity is easily visible to the eye, and in fact a considerably smaller change could be seen, or at least detected with a sensitive intensity-measuring instrument such as a microphotometer. However, the depth of the minimum changes very rapidly with separation in this region, and in view of the simplicity of the relations in this particular case, it was decided by Rayleigh to arbitrarily fix the separation $\alpha = \theta_1 = \lambda/a$ as the criterion for resolution of two diffraction patterns.

^{*} The symbol \cong will be used here to denote "is approximately equal to."

This quite arbitrary choice is known as "Rayleigh's criterion." The angle θ_1 is sometimes called the "resolving power" of the aperture a , although the ability to resolve increases as θ_1 becomes smaller. A more appropriate designation for θ_1 is the *minimum angle of resolution*.

15.7. Resolving Power of a Prism. An example of the use of this criterion for the resolving power of a rectangular aperture is found in the prism spectroscope, where the face of the prism limits the refracted beam to a rectangular section. Thus, in Fig. 15J, the minimum angle δD between two parallel beams which give rise to images on the limit of resolution is such that $\delta D = \theta_1 = \lambda/a$, where a is the width of the emerging beam. The two beams giving these images differ in wave-

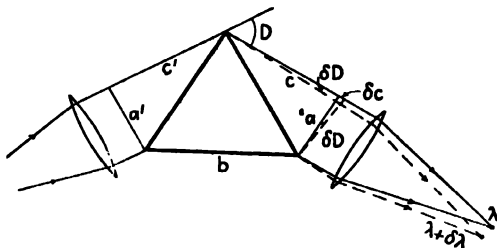


FIG. 15J. Resolving power of a prism.

length by a small increment $\delta\lambda$, which is negative because the smaller wavelengths are deviated through greater angles. The wavelength increment is more useful than the increment of angle, and hence it is customary to define the *chromatic resolving power* R of a spectroscope as the ratio $\lambda/\delta\lambda$. To evaluate this for the prism, we first note that, since any optical path between the two successive positions a' and a of the wavefront must be the same, we can write

$$c + c' = nb \quad (15n)$$

Here n is the refractive index of the prism for the wavelength λ , and b the length of the base of the prism. Now if the wavelength be decreased by $\delta\lambda$, the optical path through the base of the prism becomes $(n + \delta n)b$, and the emergent wavefront must turn through an angle $\delta D = \lambda/a$ in order that the image it forms may be just resolved. Since, from the figure, $\delta D = (\delta c)/a$, this increases the length of the upper ray by $\delta c = \lambda$. It is immaterial whether we measure δc along the rays λ or $\lambda + \delta\lambda$, because only a difference of the second order is involved. Then we have

$$c + c' + \lambda = (n + \delta n)b$$

and, subtracting Eq. 15n,

$$\lambda = b \, \delta n$$

The desired result is now obtained by dividing by $\delta\lambda$ and substituting the derivative $dn/d\lambda$ for the ratio of small increments:

$$\frac{\lambda}{\delta\lambda} = b \frac{dn}{d\lambda} \quad (15o)$$

This simple relation is useful for calculating the resolving power of prisms and can be shown to hold for two or more prisms in tandem if b is the sum of the prism bases.

15.8. Circular Aperture. The diffraction pattern formed by plane waves from a point source passing through a circular aperture is of con-

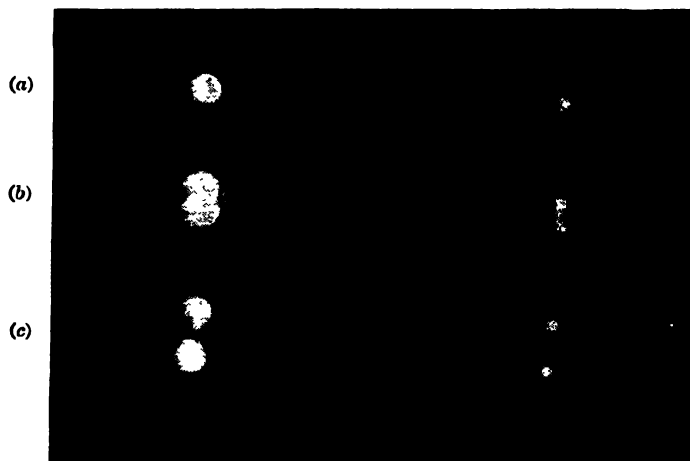


FIG. 15K. Photographs of diffraction images of point sources taken with a circular aperture. (a) One source, (b) two sources just resolved, (c) two sources completely resolved.

siderable importance as applied to the resolving power of telescopes and other optical instruments. Unfortunately it is also a problem of considerable difficulty, since it requires a double integration over the surface of the aperture similar to that mentioned in Sec. 15.5 for a rectangular aperture. The problem was first solved by Airy* in 1835, and the solution is obtained in terms of certain well-known series known as Bessel's functions. The most convenient way to express the results for our pur-

pose will be to give the numerical data obtained from calculations on these series (see Table 15I).

The diffraction pattern as illustrated in Fig. 15K(a) consists of a bright central disk, known as *Airy's disk*, surrounded by a number of fainter rings. Neither the disk nor the rings are sharply limited but shade gradually off at the edges, being separated by circles of zero intensity. The intensity distribution is very much the same as that which would be obtained with the single-slit pattern illustrated in Fig. 15E by rotating it about an axis in the direction of the light and passing through the principal maximum. The dimensions of the pattern are, however, appreciably different from those in a single-slit pattern for a slit of width equal to the diameter of the circular aperture. For the single-slit pattern, the angular separation θ of the minima from the center was found in Sec. 15.3 to be given by $\sin \theta \cong \theta = m\lambda/a$, where m is any whole number, starting with unity. The dark circles separating the bright ones in the pattern from a circular aperture may be expressed by a similar formula, if θ is now the angular semidiameter of the circle, but in this case the numbers m are not integers. Their numerical values as calculated by Lommel* are given in Table 15I. This table also

TABLE 15I

Ring	Circular aperture			Single slit	
	m	I_{\max}	I_{total}	m	I_{\max}
Central maximum.....	0	1	1	0	1
First dark.....	1.220			1.000	
Second bright.....	1.638	0.01745	0.084	1.430	0.0472
Second dark.....	2.233			2.000	
Third bright.....	2.666	0.00415	0.033	2.459	0.0168
Third dark.....	3.238			3.000	
Fourth bright.....	3.694	0.00165	0.018	3.471	0.0083
Fourth dark.....	4.241			4.000	
Fifth bright.....	4.722	0.00078	0.011	4.477	0.0050
Fifth dark.....	5.243			5.000	

includes the values of m for the maxima of the bright rings, and data on their intensities. The column headed I_{\max} gives the relative intensities of the maxima, while that headed I_{total} is the total amount of light in the ring, relative to that of the central disk. For comparison, the values of m and I_{\max} for the straight bands of the single-slit pattern are also included.

* E. V. Lommel, *Abhandl. Bayer Akad. Wiss.*, 15, 531, 1886.

15.9. Resolving Power of a Telescope. To give an idea of the linear size of the above diffraction pattern, let us calculate the radius of the first dark ring in the image formed in the focal plane of an ordinary field glass. The diameter of the objective is 4 cm and its focal length 30 cm. White light has an effective wavelength of 5.6×10^{-5} cm, so that the angular radius of this ring is $\theta = 1.220 \frac{5.6 \times 10^{-5}}{4} = 1.71 \times 10^{-5}$ rad. The linear radius is this angle multiplied by the focal length and therefore amounts to $30 \times 1.71 \times 10^{-5} = 0.000512$ cm, or almost exactly 0.005 mm. The central disk for this telescope is then 0.01 mm in diameter when the object is a point source such as a star.

Extending Rayleigh's criterion for the resolution of diffraction patterns (Sec. 15.6) to the circular aperture, two patterns are said to be resolved when the central maximum of one falls on the first dark ring of the other. The resultant pattern in this condition is shown in Fig. 15K(b). The minimum angle of resolution for a telescope is therefore

$$\theta_1 = 1.220 \frac{\lambda}{a} \quad (15p)$$

where a is the diameter of the circular aperture which limits the beam forming the primary image, or usually that of the objective. For the example chosen above, the angle calculated is just this limiting angle, so that the smallest angular separation of a double star which could be theoretically resolved by this telescope is 1.71×10^{-5} rad, or 3.52 seconds of arc. Since the minimum angle is inversely proportional to a , we see that the aperture necessary to resolve two sources 1 second apart is 3.52 times as great as in the example, or that

$$\text{Minimum angle of resolution in seconds } \theta_1 = \frac{11.1}{a} \quad (15q)$$

a being the aperture of the objective in centimeters. For the largest refracting telescope in existence, that at the Yerkes Observatory, $a = 40$ in. and $\theta_1 = 0.14$ sec. This may be compared with the minimum angle of resolution for the eye, the pupil of which has a diameter of about 3 mm. We find $\theta_1 = 47$ seconds of arc.* Actually the eye of the average person is not able to resolve objects less than about 1 minute apart, and the limit is therefore effectively determined by optical defects in the eye or by the structure of the retina.

* It might at first appear that the wavelength to be used in this calculation would be that in the vitreous humor of the eye. It is true that the dimensions of the diffraction pattern are smaller on this account, but the separation of two images is also decreased in the same proportion by refraction of the rays as they enter the eye.

With a given objective in a telescope, the angular size of the image as seen by the eye is determined by the magnification of the eyepiece. However, increasing the size of the image by increasing the power of the eyepiece does not increase the amount of detail that can be seen, since it is impossible by magnification to bring out detail which is not originally present in the primary image. Each point in an object becomes a small circular diffraction pattern or disk in the image, so that if an eyepiece of very high power is used, the image appears blurred and no greater detail is seen. Thus diffraction by the objective is the one factor that limits the resolving power of a telescope.

The diffraction pattern of a circular aperture, as well as the resolving power of a telescope, can be demonstrated by an experimental arrangement similar to that shown in Fig. 15II. The point sources at S_1 and S_2 consists of a sodium or mercury arc and a screen with several pinholes about 0.35 mm in diameter and spaced from 2 to 10 mm apart. These may be viewed with one of three small holes 1, 2, and 4 mm in diameter, mounted in front of the objective lens to show how an increasing aperture affects the resolution. Under these circumstances the intensity will not be sufficient to show anything but the central disks. In order to observe the subsidiary diffraction rings, the best source to use is the concentrated-arc lamp to be described in Sec. 21.2.

The theoretical resolving power of a telescope will be realized only if the lenses are geometrically perfect and if the magnification is at least equal to the so-called "normal" magnification (Sec. 7.14). To prove the latter statement, we note that two diffraction disks which are on the limit of resolution in the focal plane of the objective must subtend at the eye an angle of at least $\theta'_1 = 1.22 \lambda/d_e$ in order to be resolved by the eye. Here d_e represents the diameter of the eye pupil. Now according to Eq. 10k the magnification

$$M = \frac{\theta'}{\theta} = \frac{D}{d}$$

where D is the diameter of the entrance pupil (objective) and d that of the exit pupil. At the normal magnification, d is made equal to d_e , so that the normal magnification becomes

$$\frac{D}{d_e} = \frac{1.22\lambda/d_e}{1.22\lambda/d} = \frac{\theta'_1}{\theta_1}$$

Hence, if the diameter d of the exit pupil is made smaller than d_e , that of the eye pupil, we have $\theta' < \theta'_1$ and the images will cease to be resolved

by the eye even though they are resolved in the focal plane of the objective.

15.10. Brightness and Illumination of Star Images. It was proved in Sec. 7.13 that regardless of the aperture of an instrument, for magnifications up to the normal magnification the brightness of the image of an extended object remains constant and at most equal to that of the object. If the object is a point source this is no longer true, but instead the brightness increases rapidly for larger apertures. This is because all the light collected by the objective is concentrated in a diffraction pattern at its focal plane, and the area of this pattern varies inversely as the square of the diameter of the objective (Eq. 15p). Assuming normal magnification or greater, all light from the objective is admitted by the eye pupil, and the increase in brightness due to the telescope therefore equals the ratio of the area of the objective to that of the eye pupil. If the magnification is less than the normal, the eye constitutes the aperture stop and the exit pupil, and its image formed by the telescope is the entrance pupil. The ratio of their areas is the square of the magnification of the telescope, which then gives the factor by which the brightness is increased. The area of the retina illuminated remains constant, since it is determined by the diffraction pattern produced by the pupil of the eye.

The illumination of the image of a point source may be calculated by multiplying the illumination of the objective by the ratio of its area to that of the central disk of the diffraction pattern it produces, because most of the light entering the objective goes into this disk. Thus the illumination will be proportional to the area of the objective. It is chiefly for this reason that attempts are constantly being made to increase the diameter of telescope objectives. The 200-in. mirror of the Mt. Palomar telescope should permit the photography of much fainter objects than has heretofore been possible.

15.11. Resolving Power of a Microscope. In this case the same principles are applicable. The conditions are, however, different from those for a telescope, in which we were chiefly interested in the smallest permissible angular separation of two objects at a large, and usually unknown, distance. With a microscope the object is very close to the objective, and the latter subtends a large angle $2i$ at the object plane, as shown in Fig. 15L. Here we wish primarily to know the smallest distance between two points O and O' in the object which will produce images I and I' that are just resolved. Each image consists of a disk and a system of rings, as explained above, and the angular separation of two disks when they are on the limit of resolution is $\alpha = \theta_1 = 1.22\lambda/a$. When this condition holds, the wave from O' diffracted to I has zero

intensity (first dark ring), and the extreme rays $O'BI$ and $O'AI$ differ in path by 1.22λ . From the insert in Fig. 15L, we see that the $O'B$ is longer than OB or OA by $s \sin i$, and $O'A$ shorter by the same amount. The path difference of the extreme rays from O' is thus $2s \sin i$, and upon equating this to 1.22λ , we obtain

$$s = \frac{1.22\lambda}{2 \sin i} \quad (15r)$$

In this derivation, we have assumed that the points O and O' were *self-luminous* objects, such that the light given out by each has no constant phase relative to that from the other. Actually the objects used in microscopes are not self-luminous but are illuminated with light from a condenser. In this case it is impossible to have the light scattered by

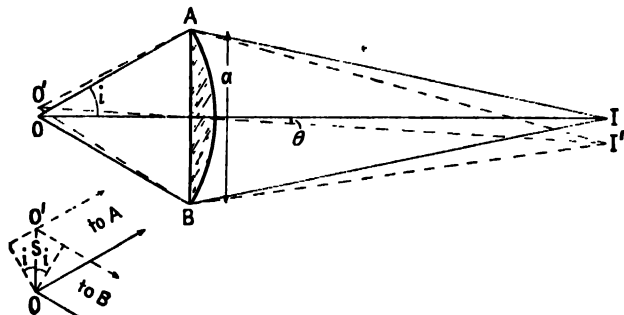


FIG. 15L. Resolving power of a microscope.

two points on the object entirely independent in phase. This greatly complicates the problem, since the resolving power is found to depend somewhat on the mode of illumination of the object. Abbe investigated this problem in detail and concluded that a good working rule for calculating the resolving power was given by Eq. 15r, omitting the factor 1.22. In microscopes of high magnifying power, the space between the object and the objective is filled with an oil. Beside decreasing the amount of light lost by reflection at the first lens surface, this increases the resolving power, because when refraction of the rays emerging from the cover glass is eliminated, the objective receives a wider cone of light from the condenser. Equation 15r must then be modified by substitution of $2ns \sin i$ for the optical path difference, where n is the refractive index of the oil. This gives

$$s = \frac{\lambda}{2n \sin i} \quad (15s)$$

The product $n \sin i$ is characteristic of a particular objective, and was called by Abbe the "numerical aperture." In practice the largest value of the numerical aperture obtainable is about 1.6. With white light of effective wavelength 5.6×10^{-5} cm, Eq. 15s gives $s = 1.8 \times 10^{-5}$ cm. The use of ultraviolet light, with its smaller value of λ , has recently been applied to still further increase the resolving power. This necessitates the use of photography in examining the image.

One of the most remarkable steps in the improvement of microscopic resolution has been the recent development of the *electron microscope*. As will be explained in Sec. 30.4, electrons behave like waves whose wavelength depends on the voltage through which they have been accelerated. For voltages between 100 and 10,000 volts, λ varies from 1.22×10^{-8} to 1.22×10^{-9} cm, *i.e.*, it lies in the region of a fraction of an angstrom unit. This is more than a thousand times smaller than for visible light. It is possible by means of electric and magnetic fields to focus the electrons emitted from, or transmitted through, the various parts of an object, and in this way details not very much larger than the wavelength of the electrons can be photographed. The numerical aperture of electron microscopes is still much smaller than that of optical instruments, but further developments in this large and growing field of *electron optics* are to be anticipated.*

15.12. Phase-contrast Microscope. This device is a modification of the ordinary microscope which is especially useful for rendering visible transparent objects that would ordinarily show little contrast. The parts of such objects which differ only in thickness or refractive index will influence the light traversing the microscope slide by altering its phase rather than its amplitude. Some success may be had in making these details visible by using variations of the ordinary mode of illumination, as for example dark-field illumination, or even by slight shifts of the focus. The truest representation of these objects, however, is obtained with the phase-contrast method, which is essentially a method of converting variations of phase on the wave front leaving the object into variations of intensity in the plane of the image. Figure 15M shows how this is done. In part (a) are shown the two essential additions to an ordinary microscope: the *phase plate* P and an *annular diaphragm* D . The latter is placed in the front focal plane of the substage condenser C , and an image of the light source is focused upon D by the concave mirror M . The object on the slide S is therefore illuminated by a hollow cone of parallel light. If there were no diffraction by objects

on the slide, this light would be focused again by the first three lenses of the objective O to form an image of D on the phase plate P . This may consist of a glass plate upon which is evaporated an annular layer of transparent material to such a thickness that it increases the optical path by one quarter of a wavelength of green light. The size of this retarding ring is such as to match the image of D . It also usually has deposited upon it a thin metallic film, to reduce its transmission.

The object on the microscope slide always has some structure which will cause diffraction of the light passing through it. For simplicity let us suppose that in Fig. 15M(b) the slide produces a diffraction pattern like the Fraunhofer pattern of a single aperture, as indicated by the

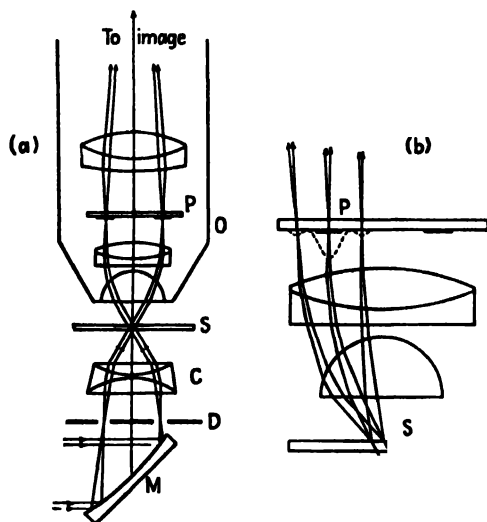


FIG. 15M. The phase-contrast microscope.

broken curve below P . Its exact form is unimportant. The light of the central maximum will undergo a phase retardation of $\pi/2$ with respect to the diffracted light. The latter is, on the average, already one-quarter vibration behind that in the central maximum, as will be seen in the vector diagram (g) of Fig. 15F. Therefore, the phase plate brings the two into phase, with a resulting large increase in the intensity at the corresponding point of the final image. The diffracting object is then rendered visible by what is known as negative or *bright contrast*. For *dark contrast*, the phase plate is made so that the direct light is advanced in phase with respect to the diffracted light. The interference at the image is then destructive, and the object is dark. For the best results, the annular portion of the phase plate is made absorbing, since otherwise

the light of the central maximum is too strong relative to the diffracted beams, and the destructive interference is not sufficiently complete. It is thus apparent that by introducing phase changes in the plane of the diffraction images, *i.e.*, in the back focal plane of the objective, an object which influences the transmitted beam only through changing its optical path may be made visible, provided that such an object produces a diffraction pattern. Although the diffraction is not as pronounced, nor as easily observable, as that produced by amplitude changes, such a pattern always exists. The above discussion also emphasizes the fact, first pointed out by Abbe, that in the complete theory of the microscopic image it is necessary to consider two stages of diffraction, one at the plane of the object, and one at the objective lens. The discussion of the last section covered only the second of these, and thus its failure to yield a quantitatively correct result becomes understandable.

Problems

1. Parallel light of wavelength 6200 \AA is incident normally on a slit 0.5 mm wide. If a lens of 50 cm focal length is mounted just behind the slit and the light focused on a screen, what will be the distance in millimeters from the center of the diffraction pattern to (a) the first minimum, (b) the first secondary maximum, and (c) the second minimum?
2. Using Eq. 15h, plot the intensity curve for the single slit in Prob. 1. Carry as far as the third minimum on each side of the center. Indicate as closely as possible the exact positions of the secondary maxima.
3. If the parallel light in Prob. 1 is produced by a slit at the focus of a lens of focal length 25 cm , how wide can this first slit be made before the details of the pattern would begin to be lost?
4. A single-slit diffraction pattern is formed with white light. Find the wavelength of light for which the third minimum coincides in position with the fourth minimum for blue light $\lambda 4500$.
5. Parallel white light is incident on a single slit 0.8 mm wide. A lens with a focal length of 80 cm brings the diffraction pattern to a focus on a screen. At a distance of 3 mm from the center of the pattern a small pinhole is made in the screen, and the transmitted light is examined with a spectroscope. What wavelengths are missing from the visible spectrum?
6. The two headlights of an automobile are 4 ft apart and 20 miles away. They are observed with a telescope having an objective 2 in. in diameter. An adjustable slit is placed in front of the objective and oriented so that its width parallel to the line between the sources can be varied. The aperture is narrowed until the two sources are barely resolved. Find its width under this condition, assuming the effective wavelength to be 5550 \AA .
7. Find the angular separation in seconds of arc of the closest double star which can be resolved with a telescope the objective of which is 6 in. in diameter.
8. The two headlights of a distant approaching automobile are 1.5 m apart. At what distance will they be resolved by the eye if the pupil is 5 mm in diameter and if the resolving power of the eye is limited by diffraction only? Assume a wavelength of 5550 \AA .

9. Calculate the chromatic resolving power of a prism having a dispersion $dn/d\lambda = -1500 \text{ cm}^{-1}$ and a base of 5 cm. Will this be adequate to resolve the two spectrum lines 5329.05 and 5329.97 Å?

10. The refractive indices of crystalline quartz are given by $n_C = 1.54190$, $n_D = 1.54425$, and $n_F = 1.54969$. The wavelengths of the C and F lines of the solar spectrum are 6563 and 4862 Å respectively. Calculate the length of base of a quartz prism which is just capable of resolving the sodium D lines, 5890 and 5896 Å.

11. The indices of refraction of calcite (ordinary ray) are $n_C = 1.65438$, $n_D = 1.65836$, and $n_F = 1.66785$. Find the base of a calcite prism required for resolution of the H_α doublet of hydrogen, $\lambda_{6562.716}$ and $\lambda_{6562.852}$.

12. Carry out the differentiation of Eq. 15*h*, equate to zero, and show that Eq. 15*j* is the resultant condition for maxima.

13. Ultraviolet light of wavelength 2537 Å has been used in photomicrography, employing quartz lenses. Assuming a numerical aperture of 0.85, what is the smallest distance apart of two points on the slide which can be resolved?

14. Violet light of wavelength 4200 Å is used with an oil-immersion microscope to resolve the lines of a diffraction grating. Find the numerical aperture required if the grating has 100,000 lines per inch.

15. Show how Eq. 15*d* leads to Eq. 15*f*. Show the steps of substituting limits and the simplification that follows.

16. The paper cone of a radio loudspeaker has a circular aperture of 12 in. Assuming it emits sound waves as in Fraunhofer diffraction, plot a graph showing the angle θ_1 (where the sound drops to the first zero minimum) as a function of frequency from 1000 vib/sec to 10,000 vib/sec. Assume the velocity of sound to be 1100 ft/sec.

17. Plot a graph showing the relative intensity at an angle of 30° from the forward direction in Prob. 16, as a function of the frequency from zero to 1000 vib/sec.

18. The parabolic reflector of a radar source, $\lambda = 3$ cm, has an aperture of 50 cm. Plot a polar graph showing the intensity as a function of the angle as far as the second zero minimum. Assume Fraunhofer diffraction.

19. Given that $R = R_0 (\sin \beta)/\beta$, where $\beta = (\pi a \sin \theta)/\lambda$, prove that $\theta = \lambda/a$ represents one-half the width of the central maximum.

20. The Fraunhofer diffraction of a single slit, reproduced twice the original size in Fig. 15*B*(c), was formed on the photographic plate in the focal plane of a lens of 1.0 m focal length. If the width of the slit was 0.095 mm, what was the wavelength of the light?

21. A small radio antenna is placed at the focus of a paraboloid of revolution, the linear diameter of the aperture being 2 m. The antenna radiates energy at 10 cm wavelength. Plot a polar diagram showing the intensity as radius vector against the angle measured from the axis of the paraboloid, for distances great enough to allow Fraunhofer diffraction considerations to be valid. What is the half-angular breadth of the central beam?

CHAPTER 16

THE DOUBLE SLIT

The interference of light from two narrow slits close together was first demonstrated by Young, and it has already been discussed in Sec. 13.2 as a simple example of the interference of two beams of light. In our discussion of the experiment, the slits were assumed to have widths not much greater than a wavelength of light, so that the central maximum

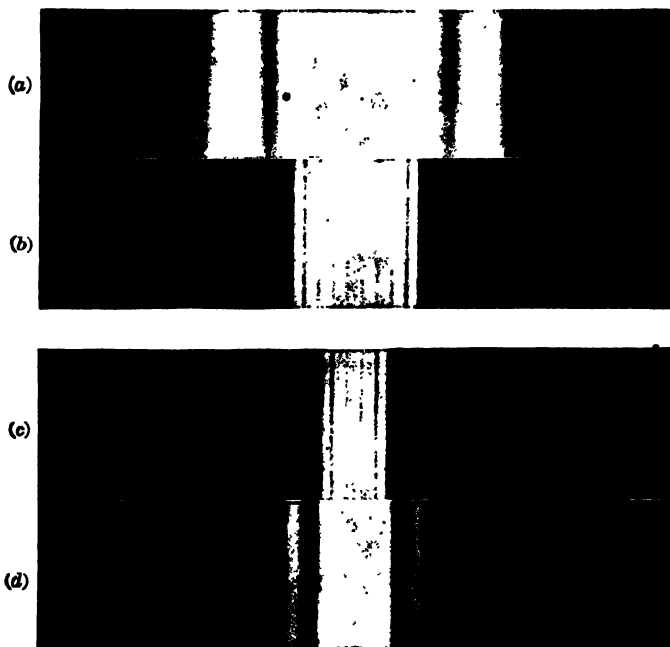


FIG. 16A. Diffraction patterns from (a) a single narrow slit, (b) two narrow slits, (c) two wider slits, (d) one wider slit.

in the diffraction pattern from each slit separately was wide enough to occupy a large angle behind the screen (see Fig. 13A). It is important to understand the modifications of the interference pattern which occur when the width of the individual slits is made greater, until it becomes comparable with the distance between them. This corresponds more nearly to the actual conditions under which the experiment is usually

performed. In this chapter we shall discuss in detail the case of *Fraunhofer diffraction* by a double slit.

16.1. Qualitative Aspects of the Pattern. In Fig. 16A(b) and (c) photographs are shown of the patterns obtained from two different double slits in which the widths of the individual slits were equal in each pair, but where the two pairs were different. Referring to Fig. 16B, which shows the experimental arrangement for photographing these patterns, the *slit width* a of each slit was greater for Fig. 16A(c) than for Fig. 16A(b), but the distance between centers $d = b + a$, or the *separation* of the slits, was the same in the two cases. In the central part of Fig. 16A(b) are seen a number of interference maxima of approximately uniform intensity, resembling the interference fringes described in Chap. 13 and shown in Fig. 13D. The intensities of these maxima are not actually

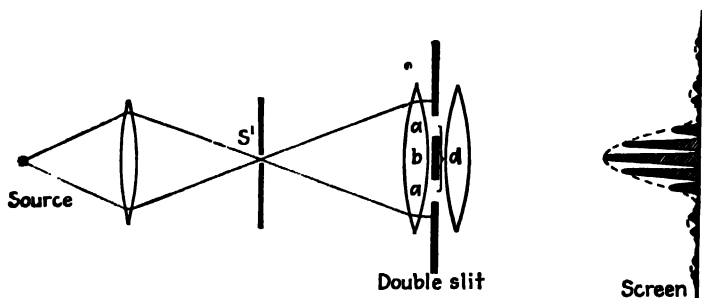


FIG. 16B. Apparatus for observing Fraunhofer diffraction from a double slit. Drawn for the case $2a = b$, that is, $d = 3a$.

constant, however, but fall off slowly to zero on either side and then reappear with low intensity two or three times before becoming too faint to observe without difficulty. The same changes are seen to occur much more rapidly in Fig. 16A(c), which was taken with greater slit widths.

16.2. Derivation of the Equation for the Intensity. Following the same procedure as that used for the single slit in Sec. 15.2, it is merely necessary to change the limits of integration in Eq. 15d to include the two portions of the wave front transmitted by the double slit. Thus if we have, as in Fig. 16B, two equal slits of width a separated by an opaque space of width b , the origin may be chosen as the center of one of the slits, and the integration is to extend from $s = -(a/2)$ to $+(a/2)$, and from $d - (a/2)$ to $d + (a/2)$, where $d = a + b$. We therefore have

$$y = \int \sin(\phi - \psi) ds = \left[\frac{s \sin \psi}{\psi} \sin \phi + \frac{s \cos \psi}{\psi} \cos \phi \right]_{-a/2}^{+a/2} + \left[\frac{s \sin \psi}{\psi} \sin \phi + \frac{s \cos \psi}{\psi} \cos \phi \right]_{d-a/2}^{d+a/2}$$

Substituting in the limits and combining terms, there results

$$y = 2a \frac{\sin \beta}{\beta} \cos \gamma \sin 2\pi \left(\frac{t}{T} - \frac{x}{\lambda} - \frac{d \sin \theta}{2\lambda} \right) \quad (16a)$$

where, as before,

$$\beta = \pi a \sin \theta$$

and where

$$\gamma = \frac{\pi}{\lambda} (a + b) \sin \theta = \frac{\pi}{\lambda} d \sin \theta \quad (16b)$$

The intensity is proportional to the square of the amplitude in Eq. 16a, so that, replacing a by R_0 as before, we have

$$I = 4R_0^2 \frac{\sin^2 \beta}{\beta^2} \cos^2 \gamma \quad (16c)$$

The factor $(\sin^2 \beta)/\beta^2$ in this equation is just that derived for the single slit of width a in the previous chapter (Eq. 15h). The second factor $\cos^2 \gamma$ is characteristic of the interference pattern produced by two beams of equal intensity and phase difference δ , as shown in Eq. 13a of Sec. 13.3. There the resultant intensity was found to be proportional to $\cos^2 (\delta/2)$, so that the expressions correspond if we put $\gamma = \delta/2$. The resultant intensity will be zero when either of the two factors is zero. For the first factor this will occur when $\beta = \pi, 2\pi, 3\pi, \dots$, and for the second factor when $\gamma = \pi/2, 3\pi/2, 5\pi/2, \dots$. That the two variables β and γ are not independent will be seen from Fig. 16C.

The difference in path from the two edges of a given slit to the screen is, as indicated, $a \sin \theta$. The corresponding phase difference is, by Eq. 11f, $(2\pi/\lambda)a \sin \theta$, which equals 2β . The path difference from any two corresponding points in the two slits is, as is illustrated for the two points at the lower edges of the slits, $d \sin \theta$, and the phase difference $\delta = (2\pi/\lambda)d \sin \theta = 2\gamma$. Therefore, in terms of the dimensions of the slits,

$$\frac{\delta}{2\beta} = \frac{\gamma}{\beta} = \frac{d}{a} \quad (16d)$$

16.3. Comparison of the Single-slit and Double-slit Patterns. It is instructive to compare the double-slit pattern with that given by a single

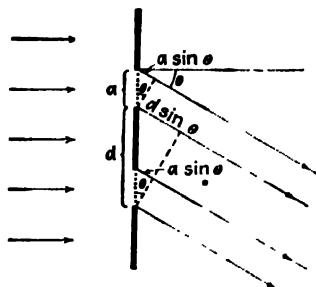


FIG. 16C. Showing path differences of parallel rays leaving a double slit.

slit of width equal to that of either of the two slits. This amounts to comparing the effect obtained with the two slits in the arrangement shown in Fig. 16B with that obtained when one of the slits is entirely blocked off with an opaque screen. If this is done, the corresponding single-slit diffraction patterns are observed, and they are related to the double-slit patterns as shown in Fig. 16A(a) and (d). It will be seen that the intensities of the interference fringes in the double-slit pattern correspond to the intensity of the single-slit pattern at any point. If one or other of the two slits is covered, we obtain exactly the same single-slit pattern in the same position, while if both slits are uncovered the pattern, instead of being a single-slit one with twice the intensity, breaks up into the narrow maxima and minima called interference fringes. The intensity at the maximum of these fringes is four times the intensity of either single-slit pattern at that point, while it is zero at the minima (see Sec. 13.4).

16.4. Distinction between Interference and Diffraction. One is quite justified in explaining the above results by saying that the light from the two slits undergoes interference to produce fringes of the type obtained with two beams, but that the intensities of these fringes are limited by the amount of light arriving at the given point on the screen by virtue of the diffraction occurring at each slit. The relative intensities in the resultant pattern as given by Eq. 16c are just those obtained by multiplying the intensity function for the interference pattern from two infinitely narrow slits of separation d (Eq. 13a) by the intensity function for diffraction from a single slit of width a (Eq. 15h). Thus, the result may be regarded as due to the joint action of interference between the rays coming from corresponding points in the two slits and of diffraction, which determines the amount of light emerging from either slit at a given angle. But diffraction is merely the result of the interference of all the secondary wavelets originating from the different elements of the wave front. Hence it is proper to say that the whole pattern is an interference pattern. It is just as correct to refer to it as a diffraction pattern, since, as we saw from the derivation of the intensity function in Sec. 16.2, it is obtained by directly summing the effects of all of the elements of the exposed part of the wave front. However, if we reserve the term *interference* for those cases in which a modification of amplitude is produced by the superposition of a finite (usually small) number of beams, and *diffraction* for those in which the amplitude is determined by an integration over the infinitesimal elements of the wave front, the double-slit pattern can be said to be due to a combination of interference and diffraction. Interference of the beams from the two slits produces the narrow maxima and minima given by the $\cos^2 \gamma$ factor, and diffraction,

represented by $(\sin^2 \beta)/\beta^2$, governs the intensities of these interference fringes. The student should not be misled by this statement into thinking that diffraction is anything other than a rather complicated case of interference.

16.5. Positions of the Maxima and Minima. Missing Orders. As shown in Sec. 16.2, the intensity will be zero wherever $\gamma = \pi/2, 3\pi/2, 5\pi/2, \dots$ and also when $\beta = \pi, 2\pi, 3\pi, \dots$. The first of these two sets are the minima for the interference pattern, and since by definition $\gamma = (\pi/\lambda)d \sin \theta$, they occur at angles θ such that

$$d \sin \theta = \frac{\lambda}{2}, \frac{3\lambda}{2}, \frac{5\lambda}{2}, \dots = \left(m + \frac{1}{2}\right) \lambda \quad \text{MINIMA} \quad (16e)$$

m being any whole number starting with zero. The second series of minima are those for the diffraction pattern, and these, since $\beta = (\pi/\lambda)a \sin \theta$, occur where

$$a \sin \theta = \lambda, 2\lambda, 3\lambda, \dots = p\lambda \quad \text{MINIMA} \quad (16f)$$

the smallest value of p being 1. The exact positions of the *maxima* are not given by any simple relation, but their approximate positions may be found by neglecting the variation of the factor $(\sin^2 \beta)/\beta^2$, a justified assumption only when the slits are very narrow and when the maxima near the center of the pattern are considered [Fig. 16A(b)]. The positions of the maxima will then be determined solely by the $\cos^2 \gamma$ factor, which has maxima for $\gamma = 0, \pi, 2\pi, \dots$, i.e., for

$$d \sin \theta = 0, \lambda, 2\lambda, 3\lambda, \dots = m\lambda \quad \text{MAXIMA} \quad (16g)$$

The whole number m represents physically the number of wavelengths in the path difference from corresponding points in the two slits (see Fig. 16C) and represents the *order* of interference.*

Figure 16D(a) is a plot of the factor $\cos^2 \gamma$, and here the values of the order, of the half-phase difference $\gamma = \delta/2$, and of the path difference are indicated for the various maxima. These are all of equal intensity and equidistant on a scale of $d \sin \theta$, or practically on a scale of θ , since when θ is small $\sin \theta \cong \theta$ and the maxima occur at angles $\theta = 0, \lambda/d, 2\lambda/d, \dots$. With a finite slit width a the variation of the factor $(\sin^2 \beta)/\beta^2$ must be taken into account. This factor alone gives just the single-slit pattern discussed in the last chapter, and is plotted in Fig. 16D(b). The complete double-slit pattern as given by Eq. 16c is the product of these two factors, and therefore is obtained by multiplying the ordinates of curve (a) by those of curve (b) and the constant $4R_0^2$. This is shown in Fig. 16D(c). The result will depend on the relative

scale of the abscissas β and γ , which in the figure are chosen so that for a given abscissa $\gamma = 3\beta$. But the relation between β and γ for a given angle θ is determined, according to Eq. 16d, by the ratio of the slit width to the slit separation. Hence if $d = 3a$, the two curves (a) and (b) are plotted to the same scale of θ . For the particular case of two slits of width a separated by an opaque space of width $b = 2a$, the curve (c), which is the product of (a) and (b), then gives the resultant pattern. The positions of the maxima in this curve are slightly different from those in curve (a) for all except the central maximum ($m = 0$), because when the ordinates near one of the maxima of curve (a) are

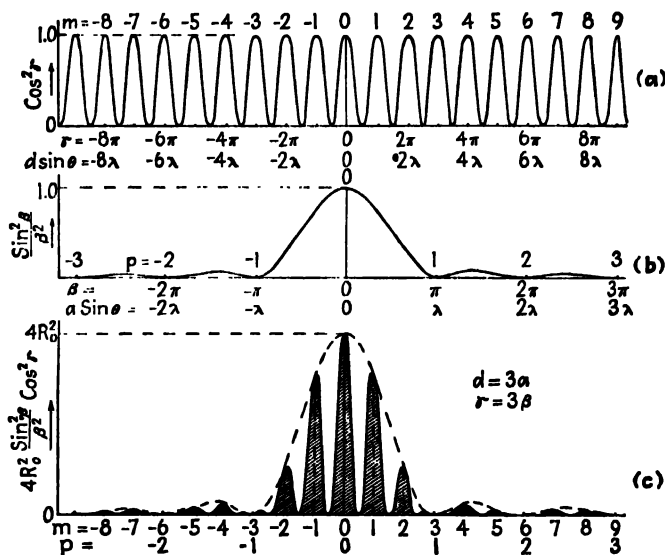


FIG. 16D. Intensity curves for a double slit where $d = 3a$.

multiplied by a factor which is decreasing or increasing, the ordinates on one side of the maximum are changed by a different amount from those of the other, and this displaces the resultant maximum slightly in the direction in which the factor is increasing. Hence the positions of the maxima in curve (c) are not exactly those given by Eq. 16g, but in most cases will be very close to them.

Let us now return to the explanation of the differences in the two patterns (b) and (c) of Fig. 16A, taken with the same slit separation d but different slit widths a . Pattern (c) was taken for the case $d = 3a$, and is seen to agree with the description given above. For pattern (b), the slit separation d was the same, giving the same spacing for the interference fringes, but the slit width a was smaller, having $d = 6a$. In

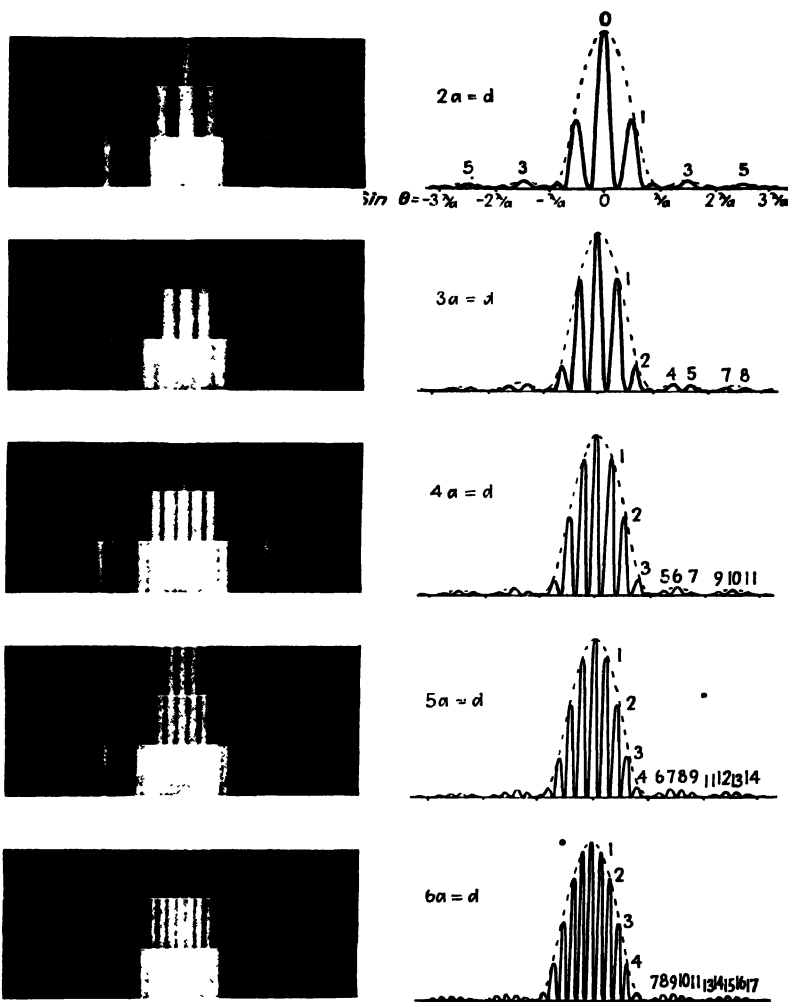


FIG. 16E. Photographs and intensity curves for double-slit diffraction patterns.

Fig. 13D, $d = 14a$. This greatly increases the scale for the single-slit pattern relative to the interference pattern, so that many interference maxima now fall within the central maximum of the diffraction pattern. Hence the effect of decreasing a , keeping d unchanged, is merely to broaden out the 'single-slit pattern, which acts as an envelope of the interference pattern as indicated by the dotted curve of Fig. 16D(c).

If the slit-width a is kept constant and the separation of the slits d is varied, the scale of the interference pattern varies, but that of the diffraction pattern remains the same. A series of photographs taken to illustrate this is shown in Fig. 16E. For each pattern three different exposures are shown, to bring out the details of the faint and the strong parts of the pattern. The maxima of the curves are labeled by the order m , and underneath the upper one is a given scale of angular positions θ . A study of these figures shows that certain orders are missing, or at least reduced to two maxima of very low intensity. These so-called *missing orders* occur where the condition for a maximum of the interference, Eq. 16g, and for a minimum of the diffraction, Eq. 16f, are both fulfilled for the same value of θ , that is for

$$d \sin \theta = m\lambda$$

$$a \sin \theta = p\lambda$$

so that

$$\frac{d}{a} = \frac{m}{p} \quad (16h)$$

Since m and p are both integers, d/a must be in the ratio of two integers in order to have missing orders. This ratio determines the orders which are missing, in such a way that when $d/a = 2$, orders 2, 4, 6, \dots are missing; when $d/a = 3$, orders 3, 6, 9, \dots are missing; etc. When $d/a = 1$ the two slits exactly join, and all orders should be missing. However, the two faint maxima into which each order is split then correspond exactly to the subsidiary maxima of a single-slit pattern of width $2a$.

Our physical picture of the cause of missing orders is as follows. Considering for example the missing order $m = +3$ in Fig. 16D(c), this point on the screen is just three wavelengths farther from the center of one slit than from the center of the other. Hence we might expect the waves from the two slits to arrive in phase and to produce a maximum. However, this point is at the same time one wavelength farther from the edge of one slit than from the other edge of that slit. Addition of the secondary wavelets from one slit gives zero intensity under these conditions. The same holds true for either slit, so that, although we may add the contributions from the two slits, both contributions are zero and must therefore give zero resultant.

16.6. Vibration Curve. The same method as that applied in Sec. 15.4 for finding the resultant amplitude graphically in the case of the single slit is applicable to the present problem. For illustration we take a double slit in which the width of each slit equals that of the opaque

space between the two, so that $d = 2a$. A photograph of this pattern appears in Fig. 16E at the top. A vector diagram of the amplitude contributions from one slit gives the arc of a circle, as before, the difference between the slopes of the tangents to the arc at the two ends being the phase difference 2β between the contributions from the two edges of the slit. Such an arc must now be drawn for each of the two slits, and the two arcs must be related in such a way that the phases (slopes of the tangents) differ for corresponding points on the two slits by 2γ , or δ . In the present case, since $d = 2a$, we must have $\gamma = \beta$ or $\delta = 4\beta$. Thus in Fig. 16F(b) showing the vibration curve for $\beta = \pi/8$, both arcs subtend an angle of $\pi/4$ ($= 2\beta$), the phase difference for the two edges of

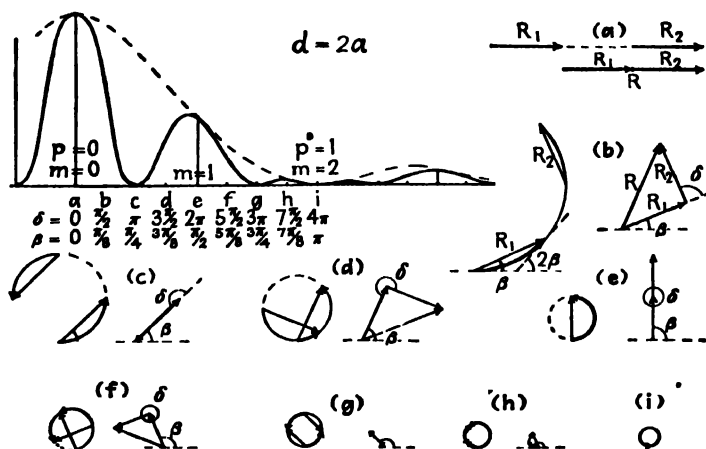


FIG. 16F. Illustrating how the intensity curve for a double slit is obtained by the graphical addition of amplitudes.

each slit, and the arcs are separated by $\pi/4$ so that corresponding points on the two arcs differ by $\pi/2$ ($= \delta$). Now the resultant contributions from the two slits are represented in amplitude and phase by the chords of these two arcs, that is by R_1 and R_2 . Diagrams (a) to (i) give the construction for the points similarly labeled on the intensity curve. The intensity, it will be remembered, is found as the square of the resultant amplitude R , which is the vector sum of R_1 and R_2 .

In the example chosen, the slits are relatively wide compared with their separation, and as the phase difference increases the curvature of the individual arcs of the vibration curve increases rapidly, so that the vectors R_1 and R_2 decrease rapidly in length. For narrower slits we obtain a greater number of interference fringes within the central diffraction maximum, because the lengths of the arcs are smaller relative to

the radius of curvature of the circle. R_1 and R_2 then decrease in length more slowly with increasing β , and the intensities of the maxima do not fall off so rapidly. In the limit where the slit width a approaches zero, R_1 and R_2 remain constant, except for the obliquity factor mentioned in Sec. 15.3, and the variation of the resultant intensity is merely due to the change in phase angle between them.

16.7. White-light Fringes. If white light is used instead of monochromatic light, the central fringe is white, and a few colored fringes are seen on either side. These are identical with the white-light fringes obtained with the Michelson interferometer (Sec. 13.13), except that in the latter case the central fringe may be black.

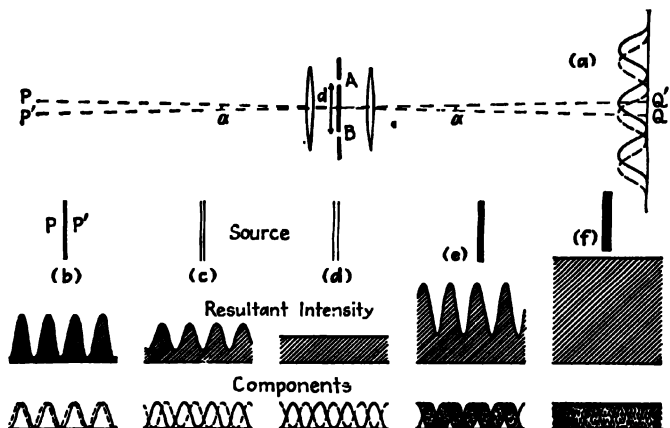


FIG. 16G. Showing the effect of a wide source on the interference fringes from a double slit.

16.8. Effect of Finite Width of Source Slit. A simplification which was made in the above treatment, and which never holds exactly in practice, is the assumption that the source slit (S' of Fig. 16B) is of negligible width. This is necessary in order that the lens shall furnish a single train of plane waves falling on the double slit. Otherwise there will be different sets of waves approaching at slightly different angles, these originating from different points in the source slit. They will produce sets of fringes which are slightly shifted with respect to each other, as illustrated in Fig. 16G(a). In the figure the interference maxima are for simplicity drawn with uniform intensity, neglecting the effects of diffraction. Let P and P' be two narrow lines acting as sources, *e.g.*, the two edges of a slit of width PP' . If the positions of the central maxima of the interference patterns produced by these are Q and Q' ,

the fringe displacement QQ' will subtend the same angle α at the double slit as do the source slits. If this angle is a small fraction of the angular separation θ_1 of the successive fringes in either pattern, the resultant intensity distribution will still resemble a true $\cos^2 \gamma$ curve, although the intensity will not fall quite to zero at the minima. The relative positions of the two patterns, and the sum of the two, in this state are illustrated in Fig. 16*G*, curves (b). Curves (c) and (d) show the effect of increasing the separation PP' . For (d) the fringes are completely out of step, and the resultant intensity shows no fluctuations whatever. At a point such as Q the maximum of one pattern then coincides with the next minimum of the other, so that the path difference $P'AQ - PAQ = \lambda/2$. In other words, P' is just a half wavelength farther from A than is P . If the intensity of one set of fringes is given by $4R^2 \cos^2 (\delta/2)$ or $2R^2(1 + \cos \delta)$, that of the other is

$$2R^2[1 + \cos (\delta + \pi)] = 2R^2(1 - \cos \delta)$$

The sum is therefore constant and equal to $4R^2$, so that the fringes entirely disappear. The condition for this disappearance of fringes is $\alpha = \theta_1/2 = \lambda/2d$. If PP' is still further increased the fringes will reappear, becoming sharp again when $\alpha = \theta_1$, then disappearing again when $\alpha = 3\theta_1/2$, etc. In general, the condition for disappearance is

$$\alpha = \frac{\lambda}{2d}, \frac{3\lambda}{2d}, \frac{5\lambda}{2d}, \dots \quad \begin{array}{l} \text{DISAPPEARANCE OF FRINGES WITH} \\ \text{DOUBLE SOURCE} \end{array} \quad (16i)$$

where α is the angle subtended by the two sources at the double slit.

Next let us consider the effect when the source, instead of consisting of two separate sources, consists of a uniformly bright strip of width PP' . Each line element of this strip will produce its own set of interference fringes, and the resultant pattern will be the sum of a large number of these, displaced by infinitesimal amounts with respect to each other. Figure 16*G*(e) illustrates this for $\alpha = \lambda/2d$, i.e., for a slit of width such that the extreme points acting alone would give complete disappearance of fringes as in (d). The resultant curve now shows strong fluctuations, and the slit must be still further widened to make the intensity uniform. The first complete disappearance will come when the range covered by the component fringes extends over a whole fringe width, instead of one-half, as above. This case is shown in Fig. 16*G*(f), for a slit of width subtending an angle $\alpha = \lambda/d$. Widening the slit still further will cause the fringes to reappear, although they never become perfectly

distinct again, with zero intensity between fringes. At $\alpha = 2\lambda/d$ they again disappear completely, and the general condition is

$$\alpha = \frac{\lambda}{d}, \frac{2\lambda}{d}, \frac{3\lambda}{d}, \dots \quad \begin{array}{l} \text{DISAPPEARANCE OF FRINGES WITH SLIT} \\ \text{SOURCE} \end{array} \quad (16j)$$

It is of practical importance, in observing double-slit fringes experimentally, to know how wide the source slit may be made in order to obtain intense fringes without seriously impairing the definition of the fringes. The exact value will depend on our criterion for clear fringes, but a good working rule is to permit a maximum discordance of the fringes of about one-quarter of that for the first disappearance. If f' is the focal length of the first lens, this corresponds to a *maximum permissible width* of the source slit

$$PP' = f'\alpha = \frac{f'\lambda}{4d} \quad (16k)$$

16.9. Michelson's Stellar Interferometer. As was shown in Sec. 15.9, the smallest angular separation that two point sources may have in order to produce images which are recognizable as separate, in the focal plane of a telescope, is $\alpha = \theta_1 = 1.22\lambda/a$. In this equation (Eq. 15p) a is the diameter of the objective of the telescope. Suppose that the objective is covered by a screen pierced with two parallel slits of separation almost equal to the diameter of the objective. A separation of $d = a/1.22$ would be a convenient value. If the telescope is now pointed at a double star and the slits are turned so as to be perpendicular to the line joining the two stars, interference fringes due to the double slit will in general be observed. However, according to Eq. 16i, if the angular separation of the two stars happens to be $\alpha = \lambda/2d$, the condition for the first disappearance, no fringes will be seen. Those from one star completely mask those from the other. Hence one could infer from the nonappearance of the fringes that the star was double with an angular separation $\lambda/2d$ or some multiple of this. (The multiples could be ruled out by direct observation without the double slit.) But this separation is only half as great as the minimum angle of resolution of the whole objective $1.22\lambda/a$, or λ/d . In this connection it is instructive to compare, as in Fig. 16H, the dimensions of the diffraction pattern due to a *rectangular* aperture of width a with the interference pattern due to two narrow slits whose separation d is equal to a . The central maximum is only half as wide in the second case. Hence it is sometimes said that the resolving power of a telescope may be increased twofold by placing a double slit over the objective. This statement needs two important qualifications, however. In the first place the stars are not "resolved"

in the sense of producing separate images, but their existence is merely inferred from the behavior of the fringes. In the second place, a partial blurring of the fringes, without complete disappearance, can be observed for separations much less than $\lambda/2d$, showing the existence of two stars, and from this point of view the minimum resolvable separation is considerably smaller than that indicated by the above statement. In practice it is about one-tenth of this.

The actual measurement of the separation of a given close double star is made by having the slit separation d adjustable. The separation is increased until the fringes first disappear; then, by measuring d , the angular separation is obtained as $\alpha = \lambda/2d$. The effective wavelength λ of the starlight must, of course, be also estimated or measured. Separations of double stars are not often determined by this method, because

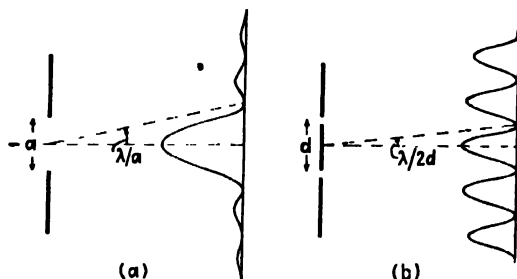


FIG. 16.17. Fraunhofer pattern from (a) a rectangular aperture, and (b) double slit of separation equal to the width of the aperture in (a).

the advantage over the direct method is not very great, and other powerful means are available which surpass the interferometric method in sensitivity (Sec. 11.6). On the other hand, the method of double-slit interference is the only one available for a direct measurement of diameter of the disk of a single star, and in 1920 this method was successfully applied by Michelson for this purpose.

From the discussion of the preceding section, it will be seen that if a source such as a star disk subtends a finite angle, disappearance of the fringes would be expected from this cause when the separation of the double slit on a telescope is made great enough. Michelson first demonstrated the practicability of this method by measuring the diameters of Jupiter's moons, which subtend an angle of about one second. The values of d for the first disappearance are only a few centimeters in this case, and the measurement could be made by a double slit of variable separation over the objective of a telescope. Owing to the fact that the source is a circular disk instead of a rectangle, a correction must be applied to the equation $\alpha = \lambda/d$ for a slit source. This correction may be found

by the same method that is used in finding the resolving power of a circular aperture, and gives the same factor. It is found that $\alpha = 1.22\lambda/d$ gives the first disappearance for a disk source. Estimating the angular diameters of the nearer fixed stars of known distance by assuming they are of the same size as the sun, one obtains angles less than 0.01 second. The separations of the double slit required to detect a disk of this size are from 20 to 40 ft. Clearly no telescope in existence could be used in the way described above for the measurement of star diameters. Another drawback would be that the fringes would be so fine that it would be difficult to separate them.

To overcome these difficulties Michelson devised his *stellar interferometer*, the principle of which is illustrated in Fig. 16I. Two plane mirrors

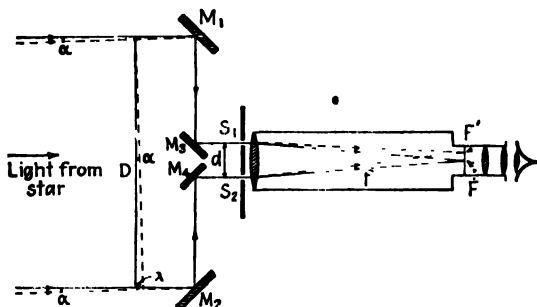


FIG. 16I. Schematic arrangement of Michelson's stellar interferometer for finding stellar diameters.

M_1 and M_2 6 in. in diameter were mounted on a long girder in front of the objective of the telescope. (In Michelson's experiment this was the 100-in. reflector at the Mt. Wilson Observatory.) These mirrors reflect the light to two other mirrors M_3 and M_4 (which were 45 in. apart), and these send two beams through the apertures of a double slit and into the telescope. Interference fringes are observed in the eyepiece, their angular separation being λ/d as for the double slit used in the ordinary way. Starting with M_1 and M_2 fairly close to the fixed mirrors in front of the telescope, they are moved out symmetrically, increasing their separation D . If at a certain value of D the fringes disappear, the angular diameter of the star $\alpha = 1.22\lambda/D$. The arrangement thus magnifies the effective resolving power of the telescope in the ratio D/d .

To prove this let us consider that for a given adjustment a point on one edge of the star disk is equidistant from M_1 and M_2 . The rays drawn as solid lines in Fig. 16I will be reflected to the double slit and will reach S_1 and S_2 exactly in phase. Thus a fringe system will be produced with its central maximum on the axis of the telescope at F .

Suppose now that the distance between M_1 and M_2 is such that a point on the opposite edge of the star disk is one wavelength farther from M_1 than from M_2 . The rays from this point (broken lines) now make an angle $\alpha = \lambda/D$ with the previous set, and produce vibrations at M_1 and M_2 which differ in phase by 2π . Because of the extreme smallness of the angle α , there will be enough diffraction at the mirrors M_1 and M_2 so that we can select a pair of rays (not obeying the ordinary law of reflection) which reach the centers of S_1 and S_2 by paths identical with those traversed by the first pair of rays. Hence these will produce vibrations at corresponding points in the two slits which differ in phase by 2π , and the resulting fringe system will be shifted one whole fringe from the previous set. The broken lines in the telescope represent the course of the light producing the central maximum F' of this second set, displaced one fringe width $f\lambda/d$ from that of the first set. The broken lines here make an angle $\theta_1 = \lambda/d$ with the solid lines, as compared to the angle $\alpha = \lambda/D$ which they make before reaching the mirrors. Intermediate points on the star disk will give interference fringes whose central maxima lie between F and F' . The net effect of all points would be to produce complete disappearance of the fringes if the star were a slit source. For a circular disk, disappearance will not be quite complete, but the mirrors would have to be moved out until $\alpha = 1.22\lambda/D$.

With a larger instrument at the Mt. Wilson Observatory, one star whose diameter was measured by this method, Betelgeuse (α Orionis), gave fringes that disappeared with the separation of the outer mirrors of 10 ft. 1 in., or 121 in. Putting this value of D in the equation and assuming $\lambda = 5700 \text{ \AA}$, we have

$$\alpha = \frac{1.22\lambda}{D} = \frac{1.22 \times 5700 \times 10^{-8}}{121 \times 2.54} = 22.7 \times 10^{-8} \text{ rad} = 0.047 \text{ sec}$$

Combining this with the distance of Betelgeuse known from its parallax, the linear diameter of the star is found to be 240 million miles. This is 278 times the diameter of the sun, so that Betelgeuse would occupy a space slightly less than the orbit of the planet Mars. Measurements have been made for several other bright stars. The smallest star measured, Arcturus, required the enormous mirror distance $D = 24 \text{ ft}$, giving an angular diameter of 0.02 second and an actual diameter 27 times as great as the sun.

16.10. Wide-angle Interference. Nothing has been said thus far as to whether there is any limit to the angle between the two interfering rays as they leave the source. Consider for example the double-slit arrangement shown in Fig. 16J(a). The source S could be a narrow

slit, but to ensure that there is no constant phase relation between the light leaving various points on it, we shall assume that it is a self-luminous object. It is found experimentally that the angle ϕ may be made fairly large without spoiling the interference fringes, provided the width of the source is made correspondingly small. Just how small it must be is seen from the fact that the path difference from the extreme edges of the source to any given point on the screen such as P must be less than $\lambda/4$ (see Sec. 16.8). Now if we call s the width of the source, the discussion given in Sec. 16.8 shows that this path difference will be $2s \sin \phi/2$. Hence, for a divergence of 60° , s cannot exceed one-quarter of a wavelength, or 1.3×10^{-5} cm for green light. If the width is made greater than this the fringes disappear completely when the path difference is λ , reappear, and then disappear again at 2λ , etc., just as in

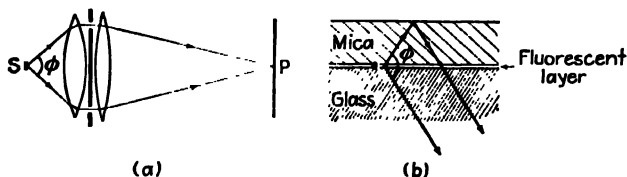


FIG. 16J. Two methods of investigating wide-angle interference.

Michelson's stellar interferometer. By using as a source an extremely thin filament, Schrödinger could still detect some interference at an angular divergence ϕ as large as 57° .

An equivalent experiment which permitted using even larger angles of divergence (up to 180°) was performed by Selenyi in 1911. The essential part of his apparatus, shown in Fig. 16J(b), was a film of a fluorescent liquid only $\frac{1}{10} \lambda$ thick contained between a thin sheet of mica and a plane glass surface. When the film is strongly illuminated it becomes a secondary source of light having a somewhat longer wavelength than the incident light (see Sec. 22.5, ahead). Interference may then be observed in a given direction between the light that comes directly from the film and that which is reflected from the outer surface of the mica. Interesting conclusions about the characteristics of the radiating atoms, in particular whether they radiate as dipoles, quadrupoles, etc., can be drawn from data on the variation of the visibility of the fringes (see footnote, page 242) with angle.

Problems

1. Make a sketch of a curve similar to that given in Fig. 16D(c) for a double slit where $d = 4a$. Only a sketch is wanted, indicating clearly the minima and central maximum.

2. Repeat Prob. 1 for the case $d = 6a$.

3. The interference pattern from a double slit having $a = 0.20$ mm and $b = 0.80$ mm is focused on a screen by means of a lens of 1 m focal length. Make a sketch as in Fig. 16D(c) of the first 12 orders on one side giving as abscissas the distances in millimeters on the screen as measured from the principal maximum. Assume $\lambda = 4000$ Å.

4. Solve Prob. 3 for the case where $a = 0.15$ mm, $b = 0.45$ mm, $f = 1.5$ m, and $\lambda = 6000$ Å.

5. Two parallel slits whose centers are 1 mm apart are each 0.25 mm wide. (a) What orders m are "missing orders"? (b) Draw the amplitude vector diagrams ("vibration curves") and resultant amplitudes for a point where the phase difference from the two slits $\delta = \pi/6$. What is the value of β for this point?

6. Solve Prob. 5 for the case where $d = 2.0$ mm, $a = 0.4$ mm, and $\delta = 2\pi/3$.

7. The first application of the principle of Michelson's stellar interferometer was in the measurement of the diameters of Jupiter's satellites. The second satellite has a diameter of 1980 miles, and at the time of measurement was 4.829×10^8 miles from the earth. At what separation of the two apertures over the telescope objective would the fringes first disappear? Assume $\lambda = 5500$ Å.

8. Calculate the relative intensity of the second-order maximum for the example shown in Fig. 16D(c). Use Eq. 16c and determine enough values near $m = 2$ to find the maximum to within 0.5 per cent.

9. Solve Prob. 8 for the fourth order, $m = 4$.

10. Parallel white light is incident on two very narrow, parallel slits for which $d = 2$ mm. A lens of focal length 2 m is used to focus the interference fringes on a screen. If a small hole is made in this screen 2 mm from the central white fringe and the light passing through is examined by a spectroscope, what wavelengths between 4000 Å and 8000 Å will be missing?

11. Solve Prob. 10 if the hole is made at 3 mm from the central white fringe.

12. Two vertical radio antennas 9 m apart are connected to the same source of power and are emitting waves with a frequency of 100 megacycles/sec. Plot a polar graph of the energy radiated in a horizontal plane.

13. As in Young's original experiment (see Sec. 13.2), interference may be produced with two circular openings. If the hole diameters are both 0.25 mm and the distance between their centers is 1.0 mm, what will be the fringe separation if the light is focused on a distant screen by a lens of 2 m focal length? Assume $\lambda = 5000$ Å.

14. Plot a graph similar to Fig. 16D for the double opening as specified in Prob. 13.

15. A double slit with $a = 0.2$ mm and $d = 0.5$ mm is mounted between two lenses as shown in Fig. 16G. If the focal length of the first lens is 1 m, how wide a source slit will just wash out the interference fringes as in diagram (f)? Assume $\lambda = 5000$ Å.

16. Given $d = a$ for a double slit. Beginning with $I = 4R_0^2 \frac{\sin^2 \beta}{\beta^2} \cos^2 \gamma$, show that the resulting intensity at the screen is the diffraction pattern given by a single slit of width $2a$.

CHAPTER 17

THE DIFFRACTION GRATING

Any arrangement which is equivalent in its action to a number of parallel equidistant slits of the same width is called a *diffraction grating*. Since the grating is a very powerful instrument for the study of spectra, we shall treat in considerable detail the intensity pattern which it produces. We shall find that the pattern is quite complex in general but

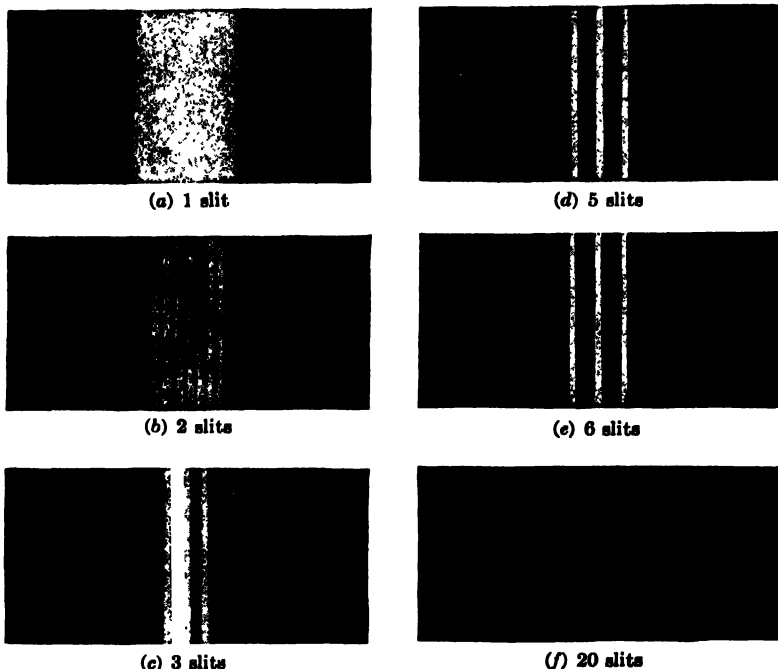


FIG. 17A. Fraunhofer diffraction patterns for gratings containing 1, 2, 3, 5, 6, and 20 equidistant slits.

that it has a number of features in common with that of the double slit treated in the last chapter. In fact, the latter may be considered as an elementary grating of only two slits. It is, however, of no use as a spectroscope, since in a practical grating many thousands of very fine slits are usually required. The reason for this becomes apparent when

we examine the difference between the pattern due to two slits and that due to many slits.

17.1. Effect of Increasing the Number of Slits. When the intensity pattern due to one, two, three, and more slits of the same width is photographed, a series of pictures like those shown in Fig. 17A(a) to (f) is obtained. The arrangement of light source, slit, lenses, and recording plate used in taking these pictures was similar to that described in previous chapters, and the light used was the blue line from a mercury arc. These patterns therefore are produced by *Fraunhofer diffraction*. In fact, it was because of Fraunhofer's original investigations of the diffraction of parallel light by gratings in 1819 that his name became associated with this type of diffraction. Fraunhofer's first gratings were made by winding fine wires around two parallel screws. Those used in preparing Fig. 17A were made by cutting narrow transparent lines in the gelatine emulsion on a photographic plate, as described in Sec. 13.2.

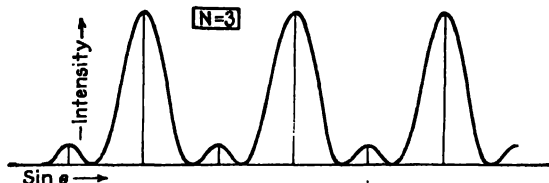


FIG. 17B. Principal and secondary maxima from a grating of three slits.

The most striking modification in the pattern as the number of slits is increased consists of a narrowing of the interference maxima. For two slits these are diffuse, having an intensity which was shown in the last chapter to vary essentially as the square of the cosine. With more slits the sharpness of these *principal maxima* increases rapidly, and in pattern (f) of the figure, with 20 slits, they have become narrow lines. This is by far the most important change introduced by increasing the number of slits. Another change which may be seen in patterns (c), (d), and (e) is the appearance of weak *secondary maxima* between the principal maxima, their number increasing with the number of slits. For three slits only one secondary maximum is present, its intensity being 11.1 per cent of the principal maximum. Figure 17B shows an intensity curve for this case, plotted according to the theoretical equation 17b given in the next section. Here the individual slits were assumed very narrow. Actually the intensities of all maxima are governed by the pattern of a single slit of width equal to that of any one of the slits used. The width of the intensity envelopes would be identical in the various patterns of Fig. 17A if the slits had been of the same width in all cases.

In fact there were slight differences in the slit widths used for some of the patterns.

17.2. General Equation for the Intensity. The required extension of the method used in Secs. 15.2 and 16.2 to derive the intensity function for the single and double slit, respectively, here consists of increasing the limits of integration to cover more than two slits. If N is the number of slits, the integral now becomes a series of N terms, each of which must be integrated within limits determined by the dimensions and position of the corresponding slit. Following our previous notation, we let a be the width of any slit, and d the constant separation between the centers of adjacent slits. The integral then becomes

$$\begin{aligned}
 y = \int \sin(\phi - \psi) ds = & \left[\frac{s \sin \psi}{\psi} \sin \phi + \frac{s \cos \psi}{\psi} \cos \phi \right]_{-\frac{a}{2}}^{+\frac{a}{2}} \\
 & + \left[\frac{s \sin \psi}{\psi} \sin \phi + \frac{s \cos \psi}{\psi} \cos \phi \right]_{d-\frac{a}{2}}^{d+\frac{a}{2}} \\
 & + \left[\frac{s \sin \psi}{\psi} \sin \phi + \frac{s \cos \psi}{\psi} \cos \phi \right]_{2d-\frac{a}{2}}^{2d+\frac{a}{2}} \\
 & + \left[\frac{s \sin \psi}{\psi} \sin \phi + \frac{s \cos \psi}{\psi} \cos \phi \right]_{3d-\frac{a}{2}}^{3d+\frac{a}{2}} \\
 & + \cdots + \left[\frac{s \sin \psi}{\psi} \sin \phi + \frac{s \cos \psi}{\psi} \cos \phi \right]_{(N-1)d-\frac{a}{2}}^{(N-1)d+\frac{a}{2}} \quad (17a)
 \end{aligned}$$

Substituting the limits, each term* will be found to contain the common factor $(a \sin \beta)/\beta$, and the remaining factors can be grouped in a series whose sum is

$$\frac{\sin N\gamma}{\sin \gamma} \sin 2\pi \left(\frac{t}{T} - \frac{x}{\lambda} - \frac{N-1}{2} \frac{d \sin \theta}{\lambda} \right)$$

* Each term, representing the contribution of the k th slit where $k = 0, 1, 2, 3, \dots, N-1$, has the form

$$a \frac{\sin \beta}{\beta} [\sin(\phi - k\delta)]$$

where $\beta = (\pi a \sin \theta)/\lambda$ and $\delta = (2\pi d \sin \theta)/\lambda$. The sum of the series is obtained from the trigonometric formula

$$\sum_{k=0}^{k=N-1} \sin(\phi - k\delta) = \frac{\sin(N\delta/2)}{\sin(\delta/2)} \sin\left(\phi - \frac{N-1}{2} \delta\right)$$

In finding the intensity we are interested only in the resultant amplitude of this expression, the square of which gives the intensity. Taking out the amplitude and squaring, the following expression is obtained

$$I = R_0^2 \frac{\sin^2 \beta}{\beta^2} \cdot \frac{\sin^2 N\gamma}{\sin^2 \gamma} \quad (17b)$$

Here β and γ have the same significance as in the previous chapter, i.e., $\beta = (\pi a \sin \theta)/\lambda$ is one-half the phase difference from opposite edges of any one slit, and $\gamma = \delta/2 = (\pi d \sin \theta)/\lambda$ is one-half the phase difference from corresponding points in any two adjacent slits.

Equation 17b should hold for any number of slits. If we put $N = 1$, it becomes $I = R_0^2 (\sin^2 \beta)/\beta^2$, which is identical with Eq. 15h for the single slit. For $N = 2$, we have

$$\begin{aligned} I &= R_0^2 \frac{\sin^2 \beta}{\beta^2} \cdot \frac{\sin^2 2\gamma}{\sin^2 \gamma} \\ &= R_0^2 \frac{\sin^2 \beta}{\beta^2} \cdot \frac{(2 \sin \gamma \cos \gamma)^2}{\sin^2 \gamma} \\ &= 4R_0^2 \frac{\sin^2 \beta}{\beta^2} \cos^2 \gamma \end{aligned}$$

agreeing with Eq. 16c for the double slit.

17.3. Principal Maxima. The factor $(\sin^2 \beta)/\beta^2$ in Eq. 17b gives the intensity distribution in the *diffraction* by a single slit, and has been discussed before (Secs. 15.2 and 16.3). The new factor $(\sin^2 N\gamma)/(\sin^2 \gamma)$ may be said to give the *interference pattern* for N slits. For $\gamma = 0, \pi, 2\pi, \dots$, it possesses maximum values equal to N^2 . Although the quotient becomes indeterminate at these values, it may be evaluated by noting that as the angle γ approaches some integral multiple of π (and therefore $N\gamma$ some other integral multiple) we have

$$\frac{\sin^2 N\gamma}{\sin^2 \gamma} = \frac{N^2 \gamma^2}{\gamma^2} = N^2$$

These maxima correspond in position to those of the double slit, neglecting the influence of the single-slit diffraction envelope, since we have for the above values of γ

$$d \sin \theta = 0, \lambda, 2\lambda, 3\lambda, \dots = m\lambda \quad \text{PRINCIPAL MAXIMA} \quad (17c)$$

This is the same as Eq. 16g for the maxima of the double-slit pattern. Multiplying the intensities of these maxima by the factor $(\sin^2 \beta)/\beta^2$, we find that their relative values are the same as those of the double slit, but that they are more intense in the ratio of the square of the

number of slits. Hence with regard to these principal maxima, we conclude that for a given slit separation d their relative intensities and positions are not changed* by increasing the number of slits, and are the same for N slits as for the two slits. The relation between β and γ in terms of slit width and slit separation, Eq. 16d, is also not changed, so that the conditions for *missing orders* (Eq. 16h) are applicable in the same form here.

17.4. Minima and Secondary Maxima. To find the minima of the function $(\sin^2 N\gamma)/(\sin^2 \gamma)$, we note that the numerator becomes zero more often than the denominator, and this occurs at the values $N\gamma = 0, \pi, 2\pi, \dots$ or, in general, $p\pi$. In the special cases when $p = 0, N, 2N, \dots$, γ will be $0, \pi, 2\pi, \dots$, so for these values the denominator will also vanish, and we have the principal maxima described above. The other values of p give zero intensity, since for these the denominator does not vanish at the same time. Hence the condition for a minimum is $\gamma = p\pi/N$, excluding those values of p for which $p = mN$, m being the order. These values of γ correspond to path differences

$$d \sin \theta = \frac{\lambda}{N}, \frac{2\lambda}{N}, \frac{3\lambda}{N}, \dots, \frac{(N-1)\lambda}{N}, \frac{(N+1)\lambda}{N}, \dots, \frac{(N+2)\lambda}{N}, \dots \quad \text{MINIMA} \quad (17d)$$

omitting the values $0, N\lambda/N, 2N\lambda/N, \dots$, for which $d \sin \theta = m\lambda$ and which according to Eq. 17c represent principal maxima. Between two adjacent principal maxima there will hence be $N - 1$ points of zero intensity. The two minima on either side of a principal maximum are separated by twice the distance of the others.

Between the other minima the intensity rises again, but the secondary maxima thus produced are of much smaller intensity than the principal maxima. Figure 17C shows a plot for six slits of the quantities $\sin^2 N\gamma$ and $\sin^2 \gamma$, and also of their quotient, which gives the intensity distribution in the interference pattern. The intensity of the principal maxima is N^2 or 36, so that the lower figure is drawn to a smaller scale. The intensities of the secondary maxima are also shown. These secondary maxima are not of equal intensity but fall off as we go out on either side of each principal maximum. Nor are they in general equally spaced, the lack of equality being due to the fact that the maxima are not quite symmetrical. This lack of symmetry is greatest for the secondary maxima immediately adjacent to the principal maxima, and is such that the

* There is a slight difference in the positions of the maxima in regions where $(\sin^2 \beta)/\beta^2$ is changing rapidly, due to the effect mentioned in Sec. 16.5.

secondary maxima are slightly shifted toward the adjacent principal maximum.

These features of the secondary maxima show a strong resemblance to those of the secondary maxima in the *single-slit* pattern. Comparison of the central part of the intensity pattern in Fig. 17C(d) with Fig. 15D for the single slit will emphasize this resemblance. As the number of slits is increased, the number of secondary maxima is also increased,

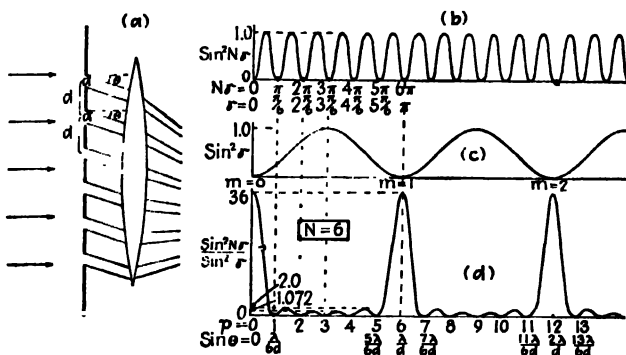


FIG. 17C. Fraunhofer diffraction by a grating of six slits, and details of the intensity pattern.

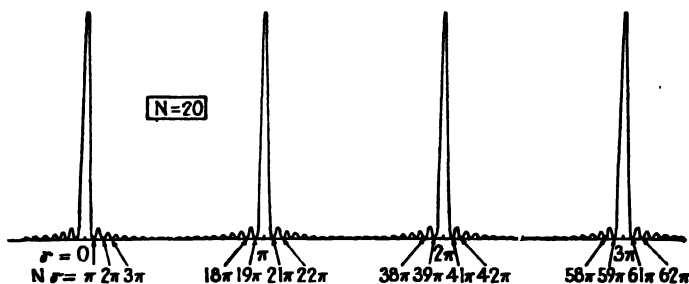


FIG. 17D. Intensity pattern for a diffraction grating of 20 slits.

since it is equal to $N - 2$. At the same time the resemblance of any principal maximum and its adjacent secondary maxima to the single-slit pattern increases. In Fig. 17D is shown the interference curve for $N = 20$, corresponding to the last photograph shown in Fig. 17A. In this case there are 18 secondary maxima between each pair of principal maxima, but only those fairly close to the principal maxima appear with an appreciable intensity. The agreement with the single-slit pattern is here practically complete, and the intensities of the secondary maxima going out from each principal maxima are very close to the values $1/22.2$, $1/61.7$, $1/121$, \dots , given in Sec. 15.3.

The reason for this correspondence will be considered in detail in Sec. 17.10, but it is worth pointing out here that one may understand it, at least for the minima adjacent to the central maximum ($m = 0$), by the following considerations. Suppose the number of slits included within a given total width of grating, such as that illustrated in Fig. 17C(a), be increased until N is very large. The slits then become very close together, and secondary wavelets are given out from each small element of the surface in practically the same manner as from an unobstructed aperture of width equal to that of the whole grating. Thus we should expect the pattern of the central maximum to approach that of a single aperture of this width as N approaches infinity. This is in fact the case, as will be seen when we examine the dimensions of the pattern in Sec. 17.8.

Even when the number of slits is small, the secondary maxima can be considered as those of a single aperture of width equal to that of the grating. The chief difference is that in this case instead of having a single principal maximum we now have several (the various orders m). Therefore the secondary maxima are somewhat more intense due to *overlapping* by those from adjacent orders. That this statement is more than qualitatively true, and that the intensities of the secondary maxima are just those calculated by summing the contributions from all orders, can be proved by a more detailed mathematical analysis of the problem. When N is small, this has a considerable effect, but as N increases, the intensities of the overlapping maxima soon become so small as to be negligible. With $N = 20$ (Fig. 17D) the intensity of the ninth secondary maximum, halfway between two orders, is only $1/N^2$ or $\frac{1}{400}$ of the principal maxima, and therefore is too weak to be experimentally observed. Even the first and second maxima are so faint relative to the principal maxima and so close to it that they are difficult to detect when N is large. In Fig. 17A they are not to be seen for $N = 20$. It should be borne in mind that the curves of Figs. 17C(d) and 17D are not yet complete representations of theory. They have still to be multiplied by the term $(\sin^2 \beta)/\beta^2$ for single-slit diffraction (Eq. 17b).

17.5. Formation of Spectra by a Grating. The secondary maxima discussed above are of little importance in the production of spectra by a many-lined grating. The principal maxima treated in Sec. 17.3 are called *spectrum lines* because when the primary source of light is a narrow slit they become sharp, bright lines on the screen. These lines will be parallel to the rulings of the grating if the slit also has this direction. For monochromatic light of wavelength λ , the angles θ at which these lines are formed are given by Eq. 17c, which is the ordinary grating equation $d \sin \theta = m\lambda$ commonly given in elementary textbooks. A

more general equation includes the possibility of light incident on the grating at any angle i . The equation then becomes

$$d(\sin i + \sin \theta) = m\lambda \quad \text{GRATING EQUATION} \quad (17e)$$

since, as will be seen from Fig. 17E, this is the path difference for light passing through adjacent slits. The figure shows the path of the light forming the maxima of order $m = 0$ (called the *central image*), and also $m = 4$ in light of a particular wavelength λ_1 . For the central image, Eq. 17e shows that $\sin \theta = -\sin i$, or $\theta = -i$. The negative sign comes from the fact that we have chosen to call i and θ positive when measured on the same side of the normal; i.e., our convention of signs is

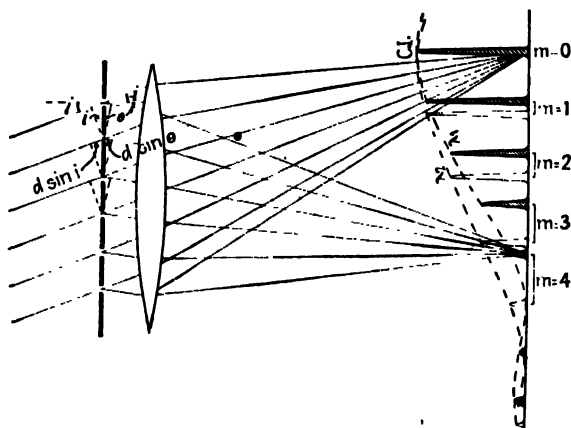


FIG. 17E. Illustrating the positions and intensities of the principal maxima from a grating, where light containing two wavelengths is incident at an angle i and diffracted at various angles θ .

such that whenever the rays used *cross over* the line normal to the grating, θ is taken as negative. Those intensity maxima which are shaded show the various orders of the wavelength λ_1 . In the case of the fourth order, for example, the path differences indicated are such that $d(\sin i + \sin \theta) = 4\lambda_1$. The intensities of the principal maxima are limited by the diffraction pattern $(\sin^2 \beta)/\beta^2$, corresponding to a single slit (broken line) and decreases to zero at the first minimum, which here coincides with the fifth order. The missing order depends on the ratio of slit separation to slit width exactly as in the double-slit case, Eq. 16h, so that here we must have $d = 5a$, and the orders $m = 5, 10, 15, \dots$ are missing orders.

Now if the source gives light of another wavelength λ_2 somewhat greater than λ_1 , the maxima of corresponding order m for this wavelength will, according to Eq. 17e, occur at larger angles θ . Since the spectrum

lines are narrow, these maxima will in general be entirely separate in each order from those of λ_1 , and we have two lines forming a *line spectrum* in each order. These spectra are indicated by brackets in the figure. Both the wavelengths will coincide, however, for the central image, because for this the path difference is zero for any wavelength. A similar set of spectra occurs on the other side of the central image, the shorter wavelength line in each order lying on the side toward the central image. Figure 17F shows actual photographs of grating spectra corresponding to the diagram of Fig. 17E. Spectrum (a) represents the spectrum in violet light of wavelength 4000 Å and (b) that in green light, 5000 Å, while (c) shows the result when these two are present at the same time.

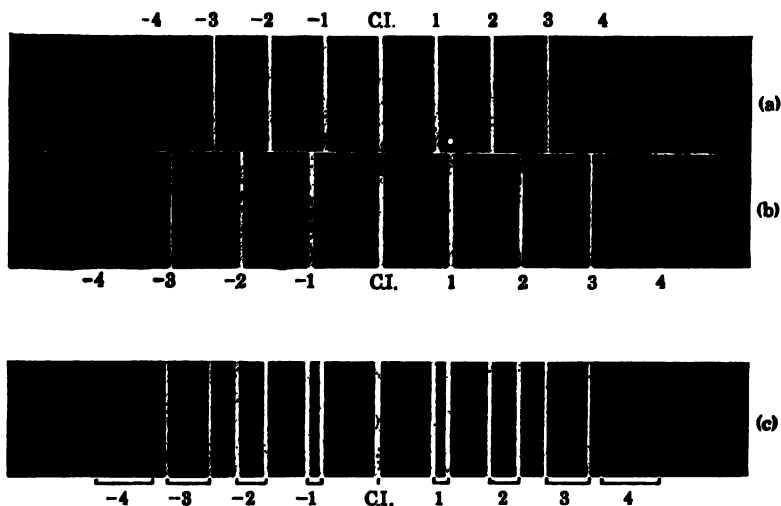


FIG. 17F. Grating spectra of two wavelengths (a) $\lambda_1 = 4000$ Å; (b) $\lambda_2 = 5000$ Å; (c) λ_1 and λ_2 together.

If the source gives white light, the central image will be white, but for the other orders each will be spread out into *continuous spectra* composed of an infinite number of adjacent images of the slit in light of the different wavelengths present. At any given point in such a continuous spectrum, the light will be very nearly monochromatic because of the narrowness of the slit images formed by the grating and lens. The result is in this respect fundamentally different from that with the double slit, where the images were broad and the spectral colors were not separated.

17.6. Dispersion. The separation of any two colors, such as λ_1 and λ_2 in Figs. 17E and 17F, increases with the order number. To express this separation the quantity frequently used is called the *angular dispersion*, which is defined as the rate of change of angle with change of wavelength.

An expression for dispersion is obtained by differentiating Eq. 17e with respect to λ , remembering that i is a constant independent of wavelength. This gives

$$\frac{d\theta}{d\lambda} = \frac{m}{d \cos \theta} \quad (17f)$$

The equation shows in the first place that for a given small wavelength difference $d\lambda$, the angular separation $d\theta$ is directly proportional to the order m . Hence the second-order spectrum is twice as wide as the first order, the third three times as wide as the first, etc. In the second place, $d\theta$ is inversely proportional to the slit separation d , which is usually referred to as the *grating space*. The smaller the grating space, the more widely spread will be the spectra. In the third place, the occurrence of $\cos \theta$ in the denominator means that in a given order m the dispersion will be smallest on the normal, where $\theta = 0$, and will increase slowly as we go out on either side of this. If θ does not become large, $\cos \theta$ will not differ much from unity, and this factor will be of little importance. If we neglect its influence, the different spectral lines in one order will differ in angle by amounts which are directly proportional to their difference in wavelength. Such a spectrum is called a *normal spectrum*, and one of the chief advantages of gratings over prism instruments is this simple linear scale for wavelengths in their spectra.

Frequently we are more interested in the linear separation of two spectrum lines on the screen or photographic plate than in their angular separation. The *linear dispersion* is obtained by multiplying Eq. 17f by the focal length f of the lens forming the spectrum. Calling l the distance along the screen, we have

$$\frac{dl}{d\lambda} = \frac{mf}{d \cos \theta} \quad (17g)$$

Here the screen is assumed to be curved into the arc of a circle of radius f with the lens at the center of curvature.

17.7. Overlapping of Orders. If the range of wavelengths is large, for instance if we observe the whole visible spectrum lying roughly between wavelengths 4000 Å and 8000 Å, considerable overlapping occurs in the higher orders. Suppose for example that one observed in the third order a certain red line of wavelength 7600 Å. The angle of diffraction for this line is given by solving for θ the expression

$$d(\sin i + \sin \theta) = 3 \times 7.6 \times 10^{-8}$$

But at the same angle θ there may occur a yellow line in the fourth order, of wavelength 5700 Å, since

$$4 \times 5.7 \times 10^{-5} = 3 \times 7.6 \times 10^{-5}$$

Similarly the blue of wavelength 4560 Å will occur in the fifth order at this same place. The general condition for the various wavelengths that can occur at a given angle θ is then

$$d(\sin i + \sin \theta) = \lambda_1 = 2\lambda_2 = 3\lambda_3 = 4\lambda_4 \quad (17h)$$

etc., where $\lambda_1, \lambda_2, \dots$ are the wavelengths in the first, second, etc., orders. For visible light there is no overlapping of the first and second orders, since with $\lambda_1 = 8000$ Å and $\lambda_2 = 4000$ Å, the red end of the first

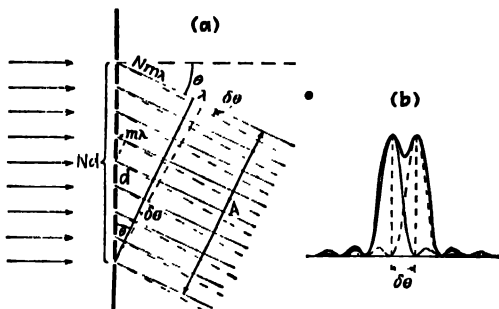


FIG. 17G. Illustrating the angular separation of two spectrum lines just resolved by a diffraction grating.

order just joins the violet end of the second. When photographic observations are made, however, these orders may extend down to 2000 Å in the ultraviolet, and there is overlapping of the first two orders. This difficulty may usually be eliminated by the use of suitable filters or absorbing screens to cut out of the incident light the overlapping wavelengths which are not desired. As an example, a piece of red glass transmitting only wavelengths longer than 6000 Å could be used in the above case to study the red line $\lambda 7600$ and lines in its vicinity without overlapping by higher orders of shorter wavelengths.

17.8. Width of the Principal Maxima. It was shown at the beginning of Sec. 17.4 that the first minima on either side of any principal maximum occur where $N\gamma = mN\pi \pm \pi$, or where $\gamma = m\pi \pm (\pi/N)$. Where $\gamma = m\pi$, we have the principal maxima, owing to the fact that the phase difference δ or 2γ in the light from corresponding points of adjacent slits is given by $2\pi m$, or a whole number of complete vibrations. However, if we change the angle enough to cause a change of $2\pi/N$ in the phase difference, reinforcement no longer occurs, but the light from the

various slits now interferes to produce zero intensity (Sec. 17.4). A phase difference of $2\pi/N$ between the maximum and the first minimum means a path difference of λ/N .

To see why this path difference causes zero intensity, consider Fig 17G(a), in which the rays leaving the grating at the angle θ form a principal maximum of order m . For these, the path difference of the rays from two adjacent slits is $m\lambda$, so that all the waves arrive in phase. The path difference of the *extreme* rays is then $Nm\lambda$, since N is always a very large number in any practical case.* Now let us change the angle of diffraction by a small amount $\delta\theta$, such that the extreme path difference increases by one wavelength and becomes $Nm\lambda + \lambda$ (rays shown by broken lines). This should correspond to the condition for zero intensity, because as is required the path difference for two adjacent slits has been increased by λ/N . It will be seen that the ray from the top of the grating is now of opposite phase from that at the center, and the effects of these two will cancel. Similarly, the ray from the next slit below the center will annul that from the next slit below the top, etc. The cancellation if continued will yield zero intensity from the whole grating, in entire analogy to the similar process considered in Sec. 15.3 for the single-slit pattern.

Thus the first zero occurs at the small angle $\delta\theta$ on each side of any principal maximum. From the figure, it is seen that

$$\delta\theta = \frac{\lambda}{A} = \frac{\lambda}{Nd \cos \theta} \quad \begin{array}{l} \text{ANGULAR HALF WIDTH OF PRINCIPAL} \\ \text{MAXIMUM} \end{array} \quad (17i)$$

It is instructive to note that this is just $1/N$ th of the separation of adjacent orders. The latter is obtained by differentiating Eq. 17e with respect to m , regarding λ as constant and m for the moment as a continuous variable. One finds

$$\frac{d\theta}{dm} = \frac{\lambda}{d \cos \theta} \cong \frac{\Delta\theta}{\Delta m}$$

For $\Delta m = 1$, the separation of orders $\Delta\theta$ is therefore N times the half width of a spectral line.

17.9. Resolving Power. The resolving power of a grating is defined just as for a prism (Sec. 15.7) by the ratio $\lambda/\delta\lambda$, where $\delta\lambda$ is the smallest wavelength difference that produces resolved images. If the Rayleigh criterion for the resolution is used, the images must be separated by the angle $\delta\theta$, given by Eq. 17i. This means that the light of wavelength $\lambda + \delta\lambda$ must form its principal maximum of order m at the same angle

* This approximation is not an essential part of the argument, since the result to be obtained is shown by Eq. 17b to be strictly accurate.

at which the first minimum for the wavelength λ occurs in the same order. Hence we may equate the extreme path differences in the two cases, and obtain

$$mN\lambda + \lambda = mN(\lambda + \delta\lambda)$$

from which it immediately follows that

$$R = \frac{\lambda}{\delta\lambda} = mN \quad (17j)$$

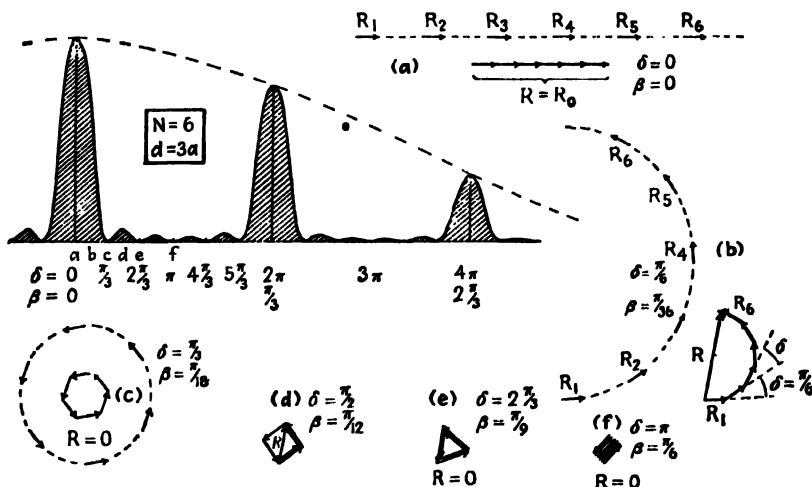
That the resolving power is proportional to the order m is to be understood from the fact that the *width* of a principal maximum, by Eq. 17i, depends on the width A of the emergent beam and does not change much with order, whereas the *separation* of two maxima of different wavelengths increases with the dispersion, which, by Eq. 17f, increases nearly in proportion to the order. In a *given order* the resolving power, by Eq. 17j, is proportional to the total number of slits N , but it is independent of their spacing d . However, at *given angles* of incidence and diffraction it is independent of N also, as can be seen by substituting in Eq. 17j the value of m from Eq. 17e:

$$R = \frac{d(\sin i + \sin \theta)}{\lambda} N = \frac{B(\sin i + \sin \theta)}{\lambda} \quad (17k)$$

Here $B = Nd$ is the total breadth of the grating. At a given i and θ , the resolving power is therefore independent of the number of lines ruled in the breadth B . A grating with fewer lines gives a higher order at these given angles, however, with the consequent overlapping, and would be of little use in practice. Nevertheless, as we shall see, this same principle is used in the echelon grating described in Sec. 17.17. It is worth noting that the maximum theoretical resolving power of a grating is obtained, according to Eq. 17k, when $i = \theta = 90^\circ$, *i.e.*, when the light is both incident and diffracted at grazing angles. Then $R = 2B/\lambda$, or twice the breadth of the grating measured in wavelengths. Actually, the intensity of the spectrum, because of the obliquity factor (Sec. 15.3), would become impractically small for the ideal grating described above, when used under these extreme conditions.

17.10. Vibration Curve. Let us now apply the method of compounding the amplitudes vectorially which was used in Sec. 16.6 for two slits and in Sec. 15.4 for one slit. The vibration curve for the contributions from the various infinitesimal elements of a single slit again forms an arc of a circle, but there are now several of these arcs in the curve, corresponding to the several slits of the grating. In Fig. 17H the diagrams corresponding to the various points (a) to (f) of the intensity

plot for six slits are shown. For the central maximum the light from all slits, and from all parts of each slit, is in phase, giving a resultant amplitude R which is N times as great as that from one slit, as shown in (a) of the figure. Halfway to the first minimum the condition is as shown in (b). For this point $\gamma = \pi/12$, so that the phase difference from corresponding points in adjacent slits δ equals $\pi/6$ (cf. Fig. 17C). This is also the angle between successive vectors in the series of six resultants R_1 to R_6 , which are the chords of the six small equal arcs. Just as for the double slit, the final resultant is obtained by compounding these vectorially, and the intensity is measured by R^2 , the square of this resultant amplitude. For the points (d), (e), and (f) the arcs them-



angle $\delta = 2\gamma$, and thus the six form part of a regular polygon. In the figure broken lines are drawn from the ends of each vector to the center O of this polygon. These lines also make the constant angle 2γ with each other. Therefore the total angle subtended at the center is

$$\phi = N\delta = N \times 2\gamma$$

We wish the relation between the resultant amplitude R and the individual ones R_n , which are given by Eq. 17l. By dividing the triangle OAB into two halves with a line from O perpendicular to R , it is seen that

$$R = 2r \sin \frac{\phi}{2}$$

where r represents OA or OB . Similarly, from the triangle OAC as split by a line perpendicular to R_1 , we obtain

$$R_n = R_1 = 2r \sin \gamma$$

Dividing this equation into the previous one, we find

$$\frac{R}{R_n} = \frac{2r \sin \frac{\phi}{2}}{2r \sin \gamma} = \frac{\sin N\gamma}{\sin \gamma}$$

FIG. 17l. Geometrical derivation of the intensity function for a grating.

When we then substitute the value of R_n from Eq. 17l, there results, for the amplitude,

$$R = R_0 \frac{\sin \beta}{\beta} \frac{\sin N\gamma}{\sin \gamma}$$

The square of this, which gives the intensity, is seen to be identical with Eq. 17b.

The vibration curve gives a rapid and accurate method of finding the intensity pattern for any number of slits, and by carrying it out for different numbers of slits, many of the features of the intensity pattern became understandable. For instance, there is the important question of the narrowness of the principal maxima. The adjacent minimum on one side is reached when the vectors first form a closed polygon, as in (c) of Fig. 17H. It is evident that this will occur for smaller values of δ the larger the number of slits, and this means that the maxima will become sharper. Also one can see at once from the diagram that for this minimum $\delta = 2\pi/N$, or $\gamma = \pi/N$, the condition stated at the beginning of Sec. 17.8. Furthermore, as the number of slits becomes large,

the polygon of vectors will rapidly approach the arc of a circle, and the analogy with the pattern due to a single aperture of width equal to that of the grating is thereby seen to be justified. Comparison of Fig. 17H with Fig. 15F for the single slit will show that for large N the diagrams for the grating will become identical with those for one slit if we replace $N\delta/2$ or $N\gamma$ by β . Since $N\gamma$ is half the phase difference from extreme slits of the grating and β half the phase difference between extreme points in an open aperture, the analogies drawn in Sec. 17.4 are seen to be justified by this correspondence.

Finally we note that if the diagrams in Fig. 17H are carried further, the first-order principal maximum occurs when the arc representing each interval d forms one complete circle. The chords under these conditions are all parallel and in the same direction as in (a), but smaller in magnitude. The second principal maximum occurs when each arc forms two turns of a circle when the resultant chords again line up. These maxima have no analogue in the pattern for a single slit. A study of the variations of these diagrams makes clear the origin of the changes in intensity of the maxima and the reason for missing orders.

17.11. Production of Ruled Gratings. Until now we have considered the characteristics of an idealized grating consisting of identical and equally spaced slits separated by opaque strips. Actual gratings used in the study of spectra are made by ruling fine grooves with a diamond point either on a plane glass surface to produce a *transmission grating* or more often on a polished metal mirror to produce a *reflection grating*. The transmission grating gives something like our idealized picture, since the grooves scatter the light and are effectively opaque, while the undisturbed parts of the surface transmit regularly and act like slits. The same is true of the reflection grating, except that here the unruled portions reflect regularly, and the grating equation 17e holds equally well for this case with the same convention of signs for i and θ .

Figure 17J shows microphotographs of the ruled surfaces of two different reflection gratings. The grating shown in (a) was ruled lightly, and the grooves are too shallow to obtain maximum brightness. That shown in (b), was a high-quality grating having 15,000 lines per inch. One or two vertical cross-rulings have been made to show more clearly the contour of the ruled surface.

Until recently, most gratings were ruled on speculum metal, a very hard alloy of copper and tin. Modern practice, however, is to rule on an evaporated layer of the softer metal aluminum. Not only does this give greater reflection in the ultraviolet, but it causes less wear on the diamond ruling point. The chief requirement for a good grating is that the lines shall be as nearly equally spaced as possible over the whole

ruled surface, which in different gratings varies from 1 to 10 in. in width. This is a difficult requirement to fulfill, and there are very few places in the world where ruling machines of precision adequate for the production of fine gratings have been constructed. After each groove has been ruled, the machine lifts the diamond point and moves the grating forward by a small rotation of the screw which drives the carriage carrying it. To have the spacing of rulings constant, the screw must be of very constant pitch, and it was not until the manufacture of a nearly perfect screw had been achieved by Rowland,* in 1882, that the problem of successfully ruling large gratings was accomplished.

In ruled gratings only the first few orders are usually employed because the overlapping of orders becomes serious above $m = 4$ or 5. Hence in

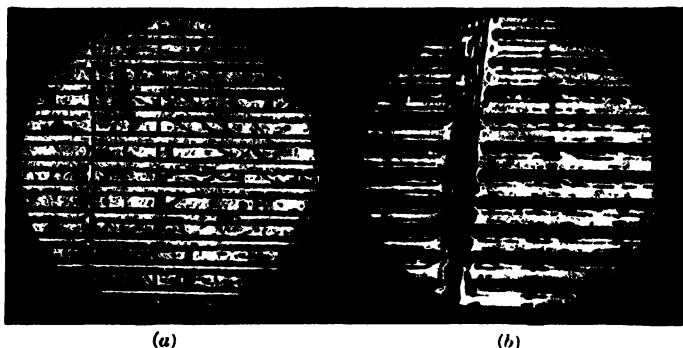


FIG. 17J. Microphotographs of the rulings on reflection gratings. (a) Light ruling. (b) Heavy ruling. (After Babcock.)

order to obtain adequate dispersion in the low orders the grating space d must be made small, according to Eq. 17f. Also, to obtain high resolving power at the same time Eq. 17j shows that N must be made as large as possible. In most gratings used in spectroscopy the lines are ruled about 15,000 to the inch (14,438 was the number given by Rowland's machine giving a grating space $d = 1.693 \times 10^{-4}$ cm). This is about equal to three wavelengths of yellow light, and according to Eq. 17e the angle of diffraction θ with normal incidence ($i = 0$) for $\lambda 5644$ in the third order is 90° . This means that it is impossible to observe wavelengths longer than this in the third order. In the second order, the limit will be, according to Eq. 17h, $\frac{2}{3} \times 5644 = 8466$ Å, in the infrared. The first order will extend to 3×5644 , or 16,932 Å. Therefore the whole

* H. A. Rowland (1848–1901). Professor of physics at the Johns Hopkins University, Baltimore. He is famous for his demonstration of the magnetic effect of a charge in motion, for his measurements of the mechanical equivalent of heat, and for his invention of the concave grating (Sec. 17.15).

visible spectrum can be observed in the first two orders, but only the violet to the yellow in the third. At the same time the dispersion will be large. The violet of the first order will, by Eq. 17e, occur at an angle θ such that

$$1.693 \times 10^{-4} \sin \theta = 1 \times 4 \times 10^{-6}$$

This gives $\theta = 13^\circ 40'$. The angle for red light of wavelength 8×10^{-6} will be $28^\circ 12'$, and the visible spectrum will be spread out between these angles. The length of the visible spectrum on the photographic plate will be the difference in angle, expressed in radians, multiplied by the focal length f of the lens which forms the spectrum. With $f = 2$ m, the first order will be 50.6 cm long, the second order 101.2 cm long, etc.

It is not the large dispersion obtainable that constitutes the chief advantage of the grating, but the narrowness of the spectrum lines, i.e., its great resolving power. A grating ruled 15,000 lines per inch for 6 in. has by Eq. 17j a resolving power in the first order of $R = \lambda/\delta\lambda = 6 \times 15,000 = 90,000$. In the green region, $\lambda = 5400$, this means that the smallest wavelength interval resolved will be $\delta\lambda = 0.06$ Å. This is only $1/100$ of the difference between the pair of yellow lines in the spectrum of sodium, which can barely be separated with a small prism spectroscope. A glass prism with a resolving power comparable to this would be impractically large, since according to Eq. 15o the length of the base b would need to be $90,000/(dn/d\lambda)$. For flint glass $dn/d\lambda$ is about 1200, so that $b = 75$ cm. Gratings are now ruled with as many as 30,000 lines to the inch for 6 in., and, although their maximum theoretical resolving power is no greater than 15,000-line gratings of the same width (Sec. 17.9), they give double the resolving power in a given order.

It was first shown by Thorp that fairly good transmission gratings could be made by taking a cast of the ruled surface with some transparent material. Such casts are called *replica gratings*, and may give satisfactory performance where the highest resolving power is not needed. Collodion or cellulose acetate, properly diluted, is poured on the grating surface and dries to a thin, tough film which can easily be detached from the master grating under water. It can then be mounted on a plane glass plate or concave mirror. Some distortion and shrinkage is involved in this process, so that the replica seldom functions as well as the master. With modern improvements in the techniques of plastics, however, it seems probable that replicas of high quality will eventually be made.

17.12. Control of the Intensity Distribution among Orders. The relative intensities of the different orders for a ruled grating do not conform to the term $(\sin^2\beta)/\beta^2$ derived for the ideal case (Eq. 17b). Obviously the light reflected from (or refracted by) the sides of the grooves will

produce important modifications. In general there will be no missing orders. The *positions* of the spectral lines are uninfluenced, however, and remain unchanged for any grating of the same grating space d . In fact, the only essential requirement for a grating is that it impress on the diffracted wave some periodic variation of either amplitude or phase. The relative intensity of different orders is then determined by the angular distribution of the light diffracted by a single element, of width d , on the grating surface. In the ideal grating this corresponds to the diffraction from a single slit. In ruled gratings it will in general be a complex factor, which in the early days of grating manufacture was considered to be largely uncontrollable. More recently, R. W. Wood has been able to produce gratings which concentrate as much as 90 per cent of the light in a single order on one side. Thus one of the chief disadvantages of gratings as compared to prisms—the presence of multiple spectra, none of which is very intense—is overcome.

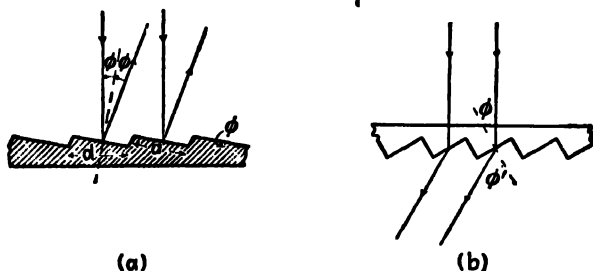


FIG. 17K. Groove forms for concentration of light in a single order (a) by reflection, (b) by transmission.

Wood's first experiments were done with gratings for the infrared, which have a large grating space so that the form of the grooves could be easily governed. These so-called "echelette" gratings had grooves with one optically flat side inclined at such an angle ϕ as to reflect the major portion of the infrared radiation toward the order that was to be bright [Fig. 17K(a)]. Of course the light from any one such face is diffracted through an appreciable angle, measured by the ratio of the wavelength to the width a of the face. When used with visible light, these gratings have interesting properties,* since one which has its "blaze" in the first order for an infrared wavelength of $6\ \mu$ will according to Eq. 17h give it in the eleventh order for the wavelength 5000 Å. Subsequent experience has shown that concentration of visible or ultraviolet light in a low order is also possible when the ruling is done on a soft metal like aluminum with a diamond point properly shaped and

oriented. A blaze can also be obtained with a transmission grating, in which case the individual rulings act like very small prisms to refract the light in the desired direction [Fig. 17K(b)]. The only successful gratings of this type have been replicas, which can be stripped from a master grating ruled at the proper angle to give the required deviation by the prismatic grooves of the replica.

17.13. Ghosts. In an actual grating the ruled lines will always deviate to some extent from the ideal of equal spacing. This gives rise to various effects, according to the nature of the ruling error. Three types may be distinguished. (1) The error is *perfectly random* in magnitude and direction. In this case the grating will give a continuous spread of light underlying the principal maxima, even when monochromatic light is used. (2) The error *continuously increases* in one direction. This can be shown to give the grating "focal properties." Parallel light after diffraction is no longer parallel, but slightly divergent or convergent. (3) The error is *periodic* across the surface of the grating. This is the most common type, since it frequently arises from defects in the driving mechanism of the ruling machine. It gives rise to "ghosts," or false lines, accompanying every principal maximum of the ideal grating. When there is only one period involved in the error, these lines are symmetrical in spacing and intensity about the principal maxima. Such ghosts are called *Rowland ghosts*, and may easily be seen in Fig. 21I(g). More troublesome, though of less frequent occurrence, are the *Lyman* ghosts*. These appear when the error involves two periods that are incommensurate with each other, or else when there is a single error of very short period. Lyman ghosts may occur very far from the principal maximum of the same wavelength.

17.14. Measurement of Wavelength with the Grating. Small gratings 1 or 2 in. wide are usually mounted on the prism table of a small spectrometer with collimator and telescope. By measuring the angles of incidence and diffraction for a given spectrum line its wavelength may be calculated from the grating formula, Eq. 17c. For this the grating space d needs to be known, and this is usually furnished with the grating. The first accurate wavelengths were determined by this method, the grating space being determined by counting the lines in a given distance with a traveling microscope. Once the absolute wavelength of a single line is known, others may be measured relative to it by using the overlapping of orders. For instance according to Eq. 17h a sodium line of wavelength 5890 Å in the third order will coincide with another line of

* Theodore Lyman (1874–). For many years director of the Physical Laboratories at Harvard University. Pioneer in the investigation of the far ultraviolet spectrum.

$\lambda = \frac{1}{2} \times 5890 = 4417 \text{ \AA}$ in the fourth order. Of course no two lines will exactly coincide in this way, but they may fall close enough together so that the small difference can be accurately corrected for. This method of comparing wavelengths is not accurate with the arrangement described above, because the telescope lens is never perfectly achromatic and the two lines will not be focused in exactly the same plane. To avoid this difficulty Rowland invented the *concave grating*, in which the focusing is done by a concave mirror, upon which the grating itself is ruled.

17.15. Concave Grating. If the grating, instead of being ruled on a plane surface, is ruled on a concave spherical mirror of metal, it will diffract and focus the light at the same time, thus doing away with the necessity of using lenses. Beside the fact that this eliminates the chro-

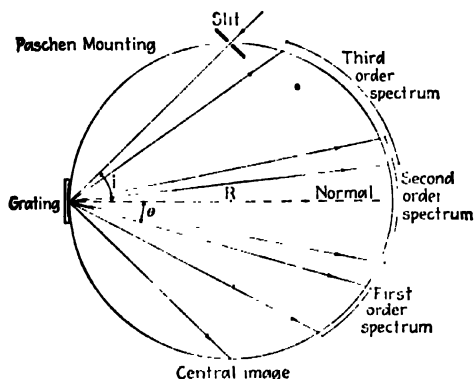


FIG. 17L. Paschen mounting for a concave grating.

matic aberration mentioned above, it has the great advantage that the grating may be used for regions of the spectrum which are not transmitted by glass lenses, such as the ultraviolet. A mathematical treatment of the action of the concave grating would be out of place here, but we may mention one of the more important results. It is found that if R is the radius of curvature of the spherical surface of the grating, a circle of diameter R (i.e., radius $r = R/2$) may be drawn tangent to the grating at its midpoint which defines the locus of points where the spectrum is in focus, provided the source slit also lies on this circle. This circle is called the *Rowland circle*, and in practically all mountings for concave gratings use is made of this condition for focus.

17.16. Mountings for Gratings. Figure 17L shows a diagram of a common form of mounting used for large concave gratings, called the *Paschen mounting*. The slit is set up on the Rowland circle, and the

light from this strikes the grating, which diffracts it into the spectra of various orders. These spectra will be in focus on the circle, and the photographic plates are mounted in a plate holder which bends them to coincide with this curve. Several orders of a spectrum can be photographed at the same time in this mounting. The ranges covered by the visible spectrum in the first three orders are indicated in Fig. 17L for the value of the grating space mentioned above. In a given order, Eq. 17f shows that the dispersion is a minimum on the normal to the grating ($\theta = 0$), and increases on both sides of this point. It is practically constant, however, for a considerable region near the normal, because here the cosine is varying slowly.

A common value for R is 21 ft, and a concave grating with this radius of curvature is called a 21-ft grating. With a grating having 15,000

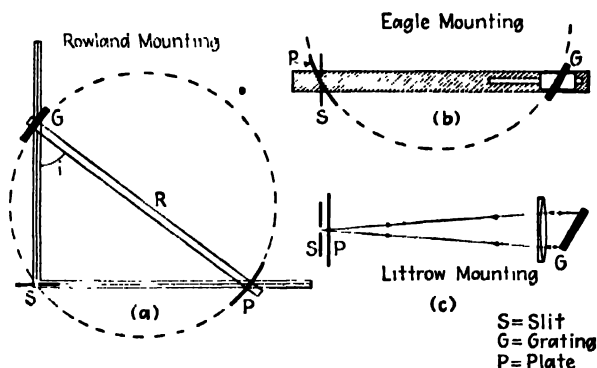


FIG. 17M. (a) Rowland mounting for a concave reflection grating. (b) Eagle mounting for a concave grating. (c) Littrow mounting for a plane reflection grating.

lines to the inch, this gives an over-all length of the second order spectrum, red to violet, of about twelve feet. A good grating having 6 in. of ruled space will produce lines sharp enough so that they may be seen as separate when only about 0.023 mm apart, which in the green means a wavelength difference of 0.03 Å (cf. Sec. 17.9).

Two other common mountings for concave gratings are the *Rowland mounting* and the *Eagle mounting*, illustrated in Fig. 17M. In the Rowland mounting the grating G and plate holder P are fixed to opposite ends of a rigid beam of length R . The two ends of this beam rest on swivel trucks which are free to move along two tracks at right angles to each other. The slit S is mounted just above the intersection of the two tracks. With this arrangement, the portion of the spectrum reaching the plate may be varied by sliding the beam one way or the other, thus varying the angle of incidence i . It will be seen that this effectively moves S around on the Rowland circle. For any setting the spectrum

will be in focus on P , and it will be nearly a normal spectrum (see Sec. 17.6) because the angle of diffraction $\theta \cong 0$. The track SP is usually graduated in wavelengths since, as may easily be shown from the grating equation, the wavelength in a given order arriving at P is proportional to the distance SP .

In the Eagle mounting the part of the spectrum is observed which is diffracted back at angles nearly equal to the angle of incidence. The slit is placed just below the plate holder. To observe different portions of the spectrum, the grating is turned about an axis perpendicular to the figure. It must then be moved along horizontal ways until P and S again lie on the Rowland circle, for the spectrum to be in focus. This mounting has the advantage of compactness, and it can easily be mounted in a small space where the temperature may be kept constant. Variations of temperature displace the spectrum lines owing to the change of grating space which results from the expansion or contraction of the grating. With a grating of speculum metal it can be shown that a change of temperature of 0.1°C shifts a line of wavelength 5000 \AA in any order by 0.013 \AA . The Eagle mounting is commonly used in *vacuum spectrographs* for the investigation of ultraviolet spectra in the region below 2000 \AA . Since air absorbs these wavelengths, the air must be pumped out of the spectrograph, and this compact mounting is convenient for the purpose. The Paschen mounting is also frequently used in vacuum spectrographs with the light incident on the grating at a practically grazing angle. The *Littrow mounting*, also shown in Fig. 17*M*, is the only common method of mounting large plane reflection gratings. In principle it is very much like the Eagle mounting, the main difference being that a large achromatic lens renders the incident light parallel and focuses the diffracted light on P , so that it acts as both collimator and telescope lenses at once.

One important drawback of the concave grating as used in the mountings described above is the presence of strong astigmatism. It is least in the Eagle mounting. This defect of the image always occurs when a concave mirror is used off axis. Here it has the consequence that each point on the slit is imaged as two lines, one located on the Rowland circle perpendicular to its plane, the other in this plane and at some distance behind the circle. If the slit is accurately perpendicular to the plane, the sharpness of the spectrum lines is not seriously impaired by astigmatism. Because of the increased length of the lines, however, some loss of intensity is involved. More serious is the fact that it is impossible to study the spectrum of different parts of a source, or to separate Fabry-Perot rings, by projecting an image on the slit of the spectrograph. For this purpose, a *stigmatic mounting* is required. The commonest of these

is the Wadsworth mounting, in which the concave grating is illuminated by parallel light. The light from the slit is rendered parallel by a large concave mirror, and the spectrum is focused at a distance of about one-half the radius of curvature of the grating.

17.17. Echelon Grating. According to Eq. 17j the resolving power of a grating is the product of the order m and the total number of lines N . In a ruled grating, high resolving power is obtained for small values of m by increasing N as much as possible. Michelson devised a type of grating which gives a very high resolving power with a fairly small value of N (20 to 40) by virtue of the large value of the order number m . This instrument is called an "echelon" because of its resemblance to a flight

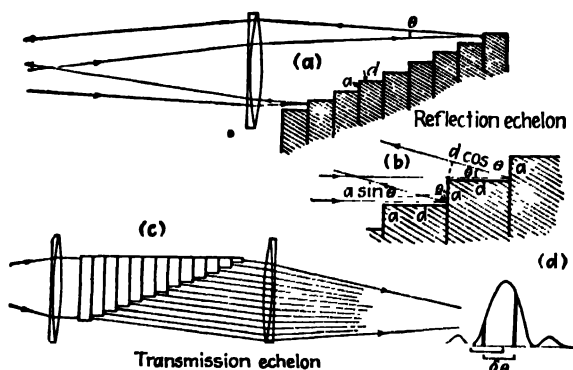


FIG. 17.N. Diagrams of reflection and transmission echelons used to obtain high resolving power.

of steps. The echelette grating mentioned in Sec. 17.12 is intermediate between the ordinary grating and the echelon. The simplest type is the *reflection echelon*, which was conceived by Michelson, but which because of technical difficulties was not successfully constructed until much later. It consists, as shown in Fig. 17N(a), of a number of rectangular plates of exactly equal thickness d , which are stacked together with a constant offset a . The plates are of quartz, and are worked into "optical contact" to insure perfect equality of the distance d between the front surfaces. These surfaces are made highly reflecting by a thin metallic coating. The offset a is usually 1 mm, so that when parallel light strikes the front surface of any plate it is diffracted only through very small angles.

Let us calculate the path difference between two rays diffracted at a small angle θ from corresponding points of two successive steps. Referring to the detailed sketch of Fig. 17N(b), the two rays will have equal paths from the dotted line to the focus of the lens, but

in reaching this line the upper one has traveled a distance $d + d \cos \theta$ while the other has traveled $a \sin \theta$. The path difference is therefore $d(1 + \cos \theta) - a \sin \theta \cong 2d - a\theta$, since θ is a very small angle. For the principal maxima this must be a whole number of wavelengths, so that the grating equation in this case becomes

$$2d - a\theta = m\lambda \quad (17m)$$

The thickness d is usually 1 cm, so the order m which is observed on the normal ($\theta = 0$) will be $m = 2d/\lambda = 2/0.00005$, or 40,000 for green light. With 40 plates, the resolving power is then

$$R = mN = 40,000 \times 40 = 1.6 \times 10^4$$

This is far greater than that reached with any ruled grating, and is rivalled only by the Fabry-Perot etalon and the Lummer-Gehrcke plate described in Chap. 14. By differentiating Eq. 17m with respect to λ and with respect to m , equations for the dispersion and angular separation of orders can be obtained. The latter quantity ($\delta\theta$ for $m = 1$) equals λ/a , a very small angle. It is just equal to the half width of the diffraction pattern due to a single face of width a . Therefore only two orders are observed with any intensity, since their intensities are limited by the diffraction pattern, as shown in Fig. 17N(d).

The successive orders are so close together in the echelon spectrum that it is of no value for examining an extended spectrum. In one order a range of only a fraction of an angstrom unit is available without overlap by the next order. Because of its high resolving power it is useful for the examination of the very fine splitting of lines called hyperfine structure (Sec. 14.9) or of the splitting of lines when the source is placed in a magnetic field (Zeeman effect). For this purpose it is necessary to isolate the line in question with an auxiliary prism or suitable filter to prevent it from being overlapped by other lines in the spectrum. Otherwise the echelon is used with slit and lens in much the same way as an ordinary plane grating in the Littrow mounting.

The *transmission echelon*, which was constructed by Michelson in 1898, bears the same relation to the reflection echelon as an ordinary transmission grating does to a reflection grating. It is made of glass plates and is used in the manner shown in Fig. 17N(c). The light emerging from successive steps has a path difference on the normal of $nd - d$, the first term being the optical path of a ray through d cm of glass and the second the path of an adjacent ray through d cm of air. The funda-

mental equation for the transmission echelon is then

$$(n - 1)d + a\theta = m\lambda \quad (17n)$$

Since $n \cong 1.5$, the order number and resolving power are only about a quarter as great as in the previous case. Otherwise the instrument is very similar to the reflection echelon.

Problems

1. Employing the method used in Sec. 15.3 in connection with the single-slit diffraction pattern, investigate the location of the secondary maxima in the multiple-slit pattern for the case $N = 6$. Exact results are not required.

2. Make a sketch of the Fraunhofer pattern obtained with eight equally spaced slits. Let the opaque intervals between slits be twice the width of each slit, i.e., $b = 2a$, or $d = 3a$.

3. Make a sketch as in Prob. 2, for 10 slits.

4. Let light of two wavelengths $\lambda 4000$ and $\lambda 4100$ fall on a diffraction grating having 5000 lines per centimeter. If a 2-m lens is used to focus the spectrum on a screen, find the distance between these two lines in centimeters on the screen (a) in the first order, (b) in the third order. Assume the focal plane to be curved and of radius 2 m.

5. Solve Prob. 4 for $\lambda 6500$ and $\lambda 6600$ if the grating has 8000 lines per centimeter.

6. If the grating in Prob. 4 is 8 cm wide, what is the smallest wavelength interval resolved in the third order at $\lambda 5000$?

7. A green spectrum line of wavelength $\lambda 5300$ is observed as a close doublet. What is the wavelength separation between these two lines if they are just resolved in the third order of a 50,000-line grating?

8. The sodium yellow line $\lambda 5893$ is a doublet, 6 Å wide. What is the minimum number of lines a grating can have to resolve this doublet in the second-order spectrum?

9. Plot the function $\sin^2 N\gamma$ against γ for $N = 4$ between $\gamma = 0$ and $\gamma = 2$. Then plot $\sin^2 \gamma$ on the same scale, and by dividing the ordinates of the first curve by those of the second, obtain the intensity curve for a grating of five very narrow slits. (NOTE: The resultant curve has maxima of intensity N^2 at $\gamma = 0, \pi$ and 2π .) Measure the intensities of the secondary maxima between each principal maximum.

10. In the adjustment of a plane grating mounted as shown in Fig. 17M(c), the incident light from the slit S makes an angle of 30° with the grating normal. Find the wavelength of the first-order light falling on the center of the plate holder directly above the slit, and calculate the dispersion. Assume a grating space of $d = 0.00025$ cm and a lens of focal length 5 m.

11. Calculate the resolving power of a reflection echelon having 20 steps, each 2 cm thick. Assume a wavelength of 4000 Å and $\theta \cong 0$ (see Sec. 17.17).

12. Find the theoretical resolving power of a transmission echelon of 30 steps, each 1 cm thick. Assume an index $n = 1.50$ and a wavelength of 5000 Å. ($dn/d\lambda = -900$ per cm.)

13. Show that, when the limits are substituted in any one term of Eq. 17a, the result simplifies to the first expression given in the footnote at the bottom of page 322.

14. Prove that in the Rowland mounting the distance SP (Fig. 17M) is proportional to the wavelength focused at P in the first order.

15. If the grating in Prob. 4 contains a total of 50,000 lines, calculate the total width of a spectrum line at $\lambda 5000$ in the first-order spectrum.

16. Five vertical radar antennas (Hertzian dipoles) are mounted in a straight horizontal line and 6 cm apart. These antennas, oscillating in synchronism, emit 3-cm waves. Plot a polar diagram of the intensity emitted in all directions in a horizontal plane. Assume Fraunhofer diffraction.

17. Find the angular width of the principal maximum of zero order in Prob. 16.

18. As with a prism, the deviation of the light by a diffraction grating depends upon the angle of incidence. Derive the condition for minimum deviation by a grating.

19. Parallel light of wavelength 5500 Å is incident upon a plane diffraction grating having 10,000 lines per centimeter, the angle of incidence being 5° . Calculate the deviations for the first orders, as measured from the direction of the zero order.

CHAPTER 18

FRESNEL DIFFRACTION

The diffraction effects obtained when either the source of light or the observing screen, or both, are at a finite distance from the diffracting aperture or obstacle come under the classification of *Fresnel diffraction*. These effects are the simplest to observe experimentally, the only apparatus required being a small source of light, the diffracting obstacle, and a screen for observation. In the Fraunhofer effects discussed in the preceding chapters, lenses were required to render the light parallel, and to focus it on the screen. Now, however, we are dealing with the more general case of divergent light which is not altered by any lenses. Since Fresnel diffraction is the easiest to observe, it was historically the first type to be investigated, although its explanation requires much more difficult mathematical theory than that necessary in treating the plane waves of Fraunhofer diffraction. In this chapter we consider only some of the simpler cases of Fresnel diffraction, which are amenable to explanation by fairly direct mathematical and graphical methods.

18.1. Shadows. One of the greatest difficulties in the early development of the wave theory of light lay in the explanation of the observed fact that light appears to travel in straight lines. Thus if we place an opaque object in the path of the light from a point source, it casts a shadow having a fairly sharp outline of the same shape as the object. It is true, however, that the edge of this shadow is not absolutely sharp and that when examined closely it shows a system of dark and light bands in the immediate neighborhood of the edge. In the days of the corpuscular theory of light, attempts were made by Grimaldi and Newton to account for such small effects as due to the deflection of the light corpuscles in passing close to the edge of the obstacle. The correct explanation in terms of the wave theory we owe to the brilliant work of Fresnel. In 1815 he showed not only that the approximately rectilinear propagation of light could be interpreted on the assumption that light is a wave motion, but also that in this way the diffraction fringes could in many cases be accounted for in detail.

To bring out the difficulty encountered in explaining shadows by the wave picture, let us consider first the passage of divergent light through an opening in a screen. In Fig. 18A the light originates from a small

pinhole S , and a certain portion MN of the divergent wave front is allowed to pass the opening. According to Huygens' principle, we may regard each point on the wave front as a source of secondary wavelets. The envelope of these at a later instant gives a divergent wave with S as its center and included between the lines SA and SB . This wave as

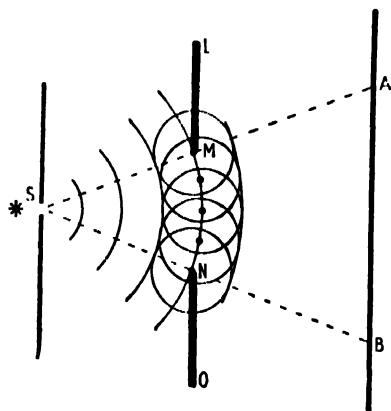


FIG. 18A. Huygens' principle applied to secondary wavelets from a narrow opening.

it advances will produce strong illumination in the region AB of the screen. But also part of each wavelet will travel into the space behind LM and NO , and hence might be expected to produce some light in the regions of the geometrical shadow outside of A and B . Common experience shows that there is actually no illumination on these parts of the screen, except in the immediate vicinity of A and B . According to Fresnel, this is to be explained by the fact that in the regions well beyond the limits of the geometrical shadow the secondary wavelets arrive with

phase relations such that they interfere destructively and produce practically complete darkness.

In two respects the simple conception of the secondary wavelets as uniform spherical waves is inadequate. In the first place, there exists another envelope of the wavelets shown in Fig. 18A, which would produce a convergent wave traveling back toward S . This *back wave* does not exist, as can be shown both experimentally and theoretically. The vibrations of the particles in the wave front MN do not set up a wave in the backward direction but merely act to annul the motions which are already possessed by the particles immediately preceding them.* The basis of the second modification has already been referred to in Sec. 15.3 as the *obliquity factor*. Theory shows that the amplitude of the secondary wavelet is not the same in all forward directions. Referring to Fig. 18B, the amplitude of the wavelet in any direction OC varies

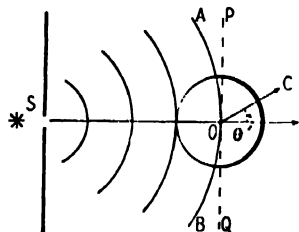


FIG. 18B. The obliquity factor for Huygens' secondary wavelets.

in proportion to $(1 + \cos \theta)$. With a value of 4 in the forward direction, the intensity drops to 1 at the tangent line PQ , and to zero in the backward direction. Thus the picture of a secondary wavelet shown in this figure is more correct, since it now takes account of this variation of amplitude with angle as well as of the absence of a wave directly backward.

18.2. Fresnel's Half-period Zones. As an example of Fresnel's approach to diffraction problems, we first consider his method of finding the effect that a slightly divergent spherical wave will produce at a point ahead of the wave. In Fig. 18C let $ABCD$ represent a spherical wave

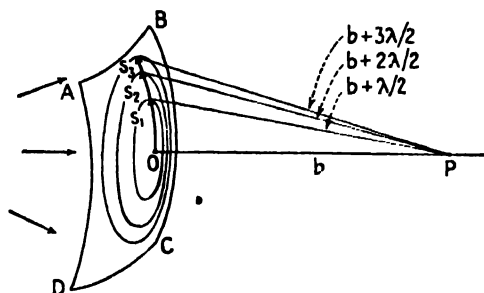


FIG. 18C. Construction of half-period zones on a spherical wave front.

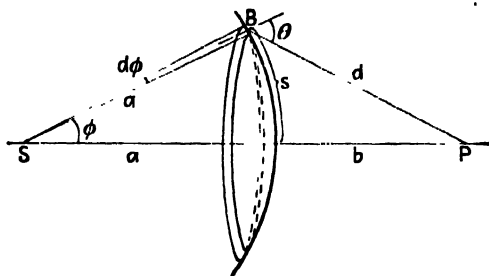


FIG. 18D. Geometry for finding the areas of Fresnel zones on a spherical wave.

front of monochromatic light traveling toward the right. Every point on this sphere may be thought of as the origin of secondary wavelets, and we wish to find the resultant effect of these at a point P . To do this, we divide the wave front into *zones* by the following construction: Around the point O , which is the foot of the perpendicular from P , we describe a series of circles whose distances from O , measured along the arc, are $s_1, s_2, s_3, \dots, s_n$ and are such that each circle is a half wavelength farther from P . If the distance $OP = b$, the circles will be at distances $b + \lambda/2, b + 2\lambda/2, b + 3\lambda/2, \dots, b + n\lambda/2$ from P .

The areas of the zones, *i.e.*, of the rings between successive circles, are practically equal. In proving this, we refer to Fig. 18D, where the wave

spreading out from a source S is shown with radius a . The area A of the segment of the sphere intercepted by the cone of half angle ϕ may be obtained by integrating the areas of elementary rings of width $a d\phi$. These are

$$dA = a d\phi \times 2\pi a \sin \phi = 2\pi a^2 \sin \phi d\phi$$

so that

$$A = 2\pi a^2 \int_0^\phi \sin \phi d\phi = 2\pi a^2 [-\cos \phi]_0^\phi$$

or

$$A = 2\pi a^2 (1 - \cos \phi) \quad (18a)$$

Now the angles ϕ_n for the borders of the zones are determined by the requirement stated above, namely that

$$d_n = b + \frac{n\lambda}{2} \quad (18b)$$

so it is necessary, in finding the areas of the zones, to evaluate $\cos \phi$ in terms of d . In the triangle SBP , the law of cosines gives

$$\cos \phi = \frac{a^2 + (a + b)^2 - d^2}{2a(a + b)}$$

Substituting this value in Eq. 18a, and putting in also d_n from Eq. 18b we find, for the area of the first n zones,

$$\begin{aligned} A &= 2\pi a^2 - \frac{2\pi a^2 \left[a^2 + (a + b)^2 - \left(b + \frac{n\lambda}{2} \right)^2 \right]}{2a(a + b)} \\ &= 2\pi a^2 - \frac{\pi a}{a + b} \left(2a^2 + 2ab - bn\lambda - \frac{n^2\lambda^2}{4} \right) \end{aligned}$$

To obtain the area A_n of the n th zone, we must then subtract the area of $n - 1$ zones, which is obtained by replacing n by $n - 1$ in the above expression for A . This yields

$$A_n = \frac{\pi a}{a + b} \left[b\lambda + \frac{(2n - 1)\lambda^2}{4} \right] \quad (18c)$$

If b is large compared to the wavelength λ , as it will always be in cases we are to consider, the term in λ^2 may be neglected. We then have

$$A_n = \frac{a}{a + b} \pi b \lambda \quad (18d)$$

This area is independent of n , and hence all zones will have approximately the same area. More accurately, the areas will increase slowly with n , according to Eq. 18c.

By Huygens' principle we now regard every point on the wave as sending out secondary wavelets in the same phase. These will reach P with different phases, since each travels a different distance. The phases of the wavelets from a given zone will not differ by more than π , and since each zone is on the average $\lambda/2$ farther from P , it is clear that the successive zones will produce resultants at P which differ by π . This statement will be examined in more detail in Sec. 18.6. The difference of half a period in the vibrations from successive zones is the origin of the name *half-period zones*. If we represent by R_n the resultant amplitude of the light from the n th zone, the successive values of R_n will have alternating signs, because changing the phase by π means reversing the direction of the amplitude vector. Calling the resultant amplitude due to the whole wave R , it may be then written as the sum of the series

$$R = R_1 - R_2 + R_3 - R_4 + \cdots + (-1)^{n-1} R_n \quad (18e)$$

There are three factors which determine the magnitudes of the successive terms in this series: First, because the area of each zone determines the number of wavelets it contributes, the terms should be approximately equal but should increase slowly; second, since the amplitude decreases inversely with the average distance of a zone d_n , the magnitudes of the terms are reduced by an amount which increases with n ; and third, because of the increasing obliquity, their magnitudes should decrease. Thus we may express the amplitude due to the n th zone as

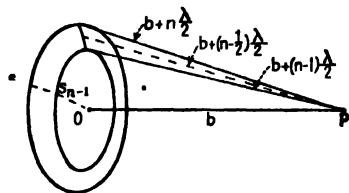


FIG. 18E. Geometry of a single half-period zone.

$$R_n = \text{const.} \times \frac{A_n}{d_n} (1 + \cos \theta) \quad (18f)$$

where θ is the angle at which the light reaching P leaves a zone. It appears in the form shown because of the obliquity factor mentioned in the preceding section.

It will be shown that the second factor mentioned above just annuls the first. From Eq. 18c the exact area of the n th zone is

$$a_n = \pi \lambda \frac{a}{a+b} \left[b + \left(n - \frac{1}{2} \right) \frac{\lambda}{2} \right] \quad (18g)$$

while the mean distance d_n of this zone from P can be seen from Fig. 18E to be

$$d_n = b + \left(n - \frac{1}{2} \right) \frac{\lambda}{2}$$

The ratio A_n/d_n is thus a constant, independent of n . Therefore we have left only the effect of the obliquity factor $1 + \cos \theta$, which causes the successive terms in Eq. 18e to decrease very slowly. The decrease is least slow at first, because of the rapid change of θ with n , but the amplitudes soon become more nearly equal.

With this knowledge of the variation in magnitude of the terms, we may evaluate the sum of the series by grouping its terms in the following two ways. Supposing n to be odd,

$$\begin{aligned} R &= \frac{R_1}{2} + \left(\frac{R_1}{2} - R_2 + \frac{R_3}{2} \right) + \left(\frac{R_3}{2} - R_4 + \frac{R_5}{2} \right) + \cdots + \frac{R_n}{2} \\ &= R_1 - \frac{R_2}{2} - \left(\frac{R_2}{2} - R_3 + \frac{R_4}{2} \right) - \left(\frac{R_4}{2} - R_5 + \frac{R_6}{2} \right) \\ &\quad \quad \quad - \cdots - \frac{R_{n-1}}{2} + R_n \end{aligned}$$

Now since the amplitudes R_1, R_2, \dots do not decrease at a uniform rate, each amplitude is smaller than the arithmetic mean of the preceding and following ones. Therefore the quantities in parentheses in the above equations are all positive, and the following inequalities must hold:

$$\frac{R_1}{2} + \frac{R_n}{2} < R < R_1 - \frac{R_2}{2} - \frac{R_{n-1}}{2} + R_n$$

Now the amplitudes for any two adjacent zones are very nearly equal, so that we may equate R_1 to R_2 , and R_{n-1} to R_n . This gives

$$\frac{R_1}{2} + \frac{R_n}{2} = R = \frac{R_1}{2} + \frac{R_n}{2} \quad (18h)$$

If n is taken to be even, the same result is obtained. We conclude that the resultant amplitude at P due to n zones is half the sum of the amplitudes contributed by the first and last zones. If we allow n to become large enough so that the whole spherical wave is divided into zones, θ approaches 180° for the last zone. Therefore the obliquity factor causes R_n to become negligible, and the amplitude due to the whole wave is just half that due to the first zone acting alone.

A very clear explanation of this result can be obtained by a simple graphical construction, representing the amplitudes by vectors as in Fig. 18F. AB is the resultant amplitude from the first zone (assumed positive), CD the smaller negative amplitude from the second, EF the

still smaller positive amplitude from the third, etc. In Fig. 18*F*(a) these are drawn separate to show their magnitudes and relative positions when added, while in (b) their true positions are shown along the same line. The resultant of the first two zones is the small vector AD ; of the first three, the large vector AF ; of the first four, the small vector AH ; etc. It will be clear from the diagram (c) that as n becomes very large $R = R_1/2$.

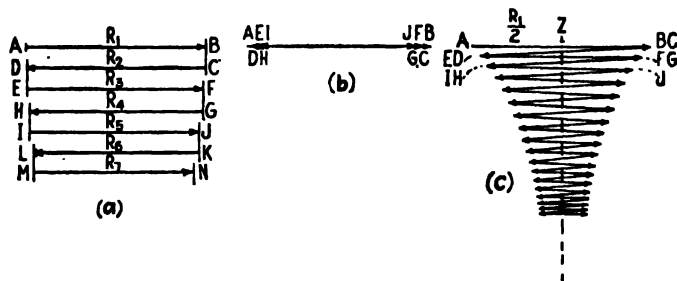


FIG. 18*F*. Amplitude diagrams for half-period zones.

18.3. Diffraction by a Circular Aperture. Let us examine the effect upon the intensity at P (Fig. 18*G*) of blocking off the wave by a screen pierced by a small circular aperture as shown in Fig. 18*G*. If the hole has a radius r equal to the distance s_1 to the outer edge of the first half-period zone,* the amplitude will be AB of Fig. 18*F*, and this is twice the amplitude due to the unscreened wave. Thus the intensity at P is four times as great as if the screen were absent. Increasing the radius of the hole until it includes the first two zones, the amplitude is AD , or practically zero. The intensity has now fallen to almost zero by increasing the size of the hole. A further increase of r will cause the intensity to pass through maxima and minima each time the number of zones included becomes odd or even.

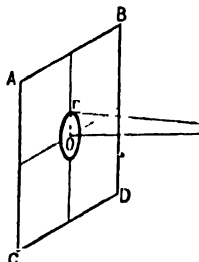


FIG. 18*G*. Geometry for light passing through a circular opening.

The same effect is produced by moving the point of observation P continuously toward or away from the aperture along the perpendicular. This varies the size of the zones, so that if P is originally at a position such that $Pr - PO$ of Fig. 18*G* is $\lambda/2$ (one zone included), moving P toward the screen will increase this path difference to $2\lambda/2$ (two zones),

* We are here assuming that the radius of curvature of the wave striking the screen is large, so that distances measured along the chord may be taken as equal to those measured along the arc.

$3\lambda/2$ (three zones), etc. We thus have maxima and minima along the axis of the aperture.

The above considerations give no information about the intensity at points off the axis. A mathematical investigation, which we shall not discuss because of its complexity, shows that P is surrounded by a system of circular diffraction fringes. Several photographs of these fringes are illustrated in Fig. 18H. These were taken by placing a photographic plate some distance behind circular holes of various sizes, illuminated by

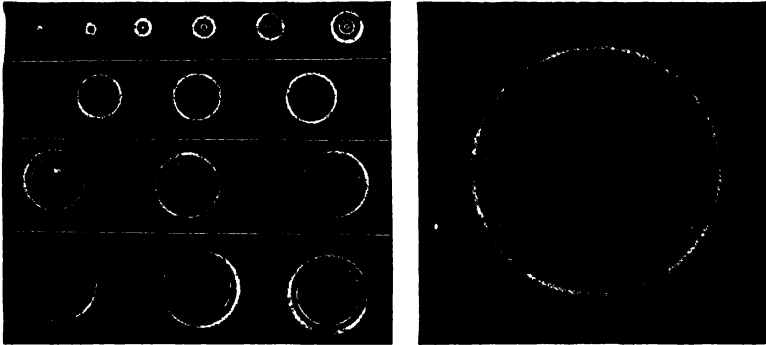


FIG. 18H. Diffraction of light by a small circular opening. (Original photographs by Huford.)

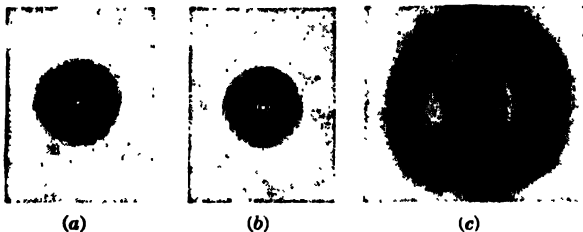


FIG. 18I. Diffraction by a circular obstacle. (a) and (b) Point source. (c) A negative of Woodrow Wilson as a source. (After Huford.)

monochromatic light from a distant point source. Starting at the upper left of the figures, the holes were of such sizes as to expose one, two, three, etc., zones. The alternation of the center of the pattern from bright to dark illustrates the result obtained above. The large pattern on the right was produced by an aperture containing 71 zones.

18.4. Diffraction by a Circular Obstacle. When the hole is replaced by a circular disk, Fresnel's method leads to the surprising conclusion that there should be a bright spot in the center of the shadow. For a treatment of this case, it is convenient to start constructing the zones at the edge of the disk. If, in Fig. 18G, $Pr = d$, the outer edge of the first zone will be $d + (\lambda/2)$ from P , of the second $d + (2\lambda/2)$, etc. The sum

of the series representing the amplitudes from the zones in this case is, as before, half the amplitude from the first zone, so the intensity at P is practically equal to that produced by the unobstructed wave. This holds only for a point on the axis, however, and off the axis the intensity is small, showing faint concentric rings. In Fig. 18I(a) and (b), which show photographs of the bright spot, these rings are unduly strengthened relative to the spot by overexposure. In (c) the source, instead of being a point, was a photographic negative of a portrait of Woodrow Wilson on a transparent plate, illuminated from behind. The disk acts like a rather crude lens in forming an image, since for every point in the object there is a corresponding bright spot in the image.

The complete investigation of diffraction by a circular obstacle shows that, besides the spot and faint rings in the shadow, there are bright circular fringes bordering the outside of the shadow. These are similar in origin to the diffraction fringes from a straight edge to be investigated in Sec. 18.10.

The bright spot in the center of the shadow of a 1-cent piece may easily be seen by examining the region of the shadow produced by an arc light several meters away, preferably using a magnifying glass.

18.5. Zone Plate. This is a special screen designed to block off the light from every other half-period zone. The result is to remove either all the positive terms in Eq. 18e or all the negative terms. In either case

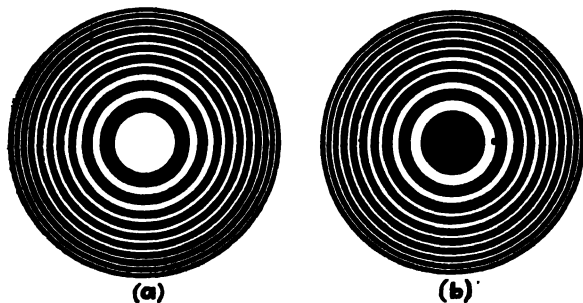


FIG. 18J. Zone plates.

the amplitude at P (Fig. 18C) will be increased to many times its value in the above cases. A zone plate can easily be made in practice by drawing concentric circles on white paper, with radii proportional to the square roots of whole numbers (see Fig. 18J). Every other zone is then blackened, and the result is photographed on a reduced scale. The negative, when held in the light from a distant point source, produces a large

intensity at a point on its axis at a distance corresponding to the size of the zones and the wavelength of the light used. The relation between these quantities is easily obtained if we assume the source at infinity, giving a plane wave and plane zones. In Fig. 18C we then have a series of right triangles and may write, for any one of them,

$$b^2 + s_n^2 = \left(b + \frac{n\lambda}{2}\right)^2$$

This yields, for the exact radii of the zones,

$$s_n = \sqrt{nb\lambda + \frac{n^2}{4}\lambda^2}$$

or practically, since we assume λ small compared to b ,

$$s_n = \sqrt{nb\lambda}$$

This justifies the statement made above that the radii should be proportional to \sqrt{n} .

The bright spot produced by a zone plate is so intense that the plate acts much like a lens. Thus suppose that the first 10 odd zones are exposed, as in the zone plate of Fig. 18J(a). This leaves the amplitudes $R_1, R_3, R_5, \dots, R_{19}$ [see Fig. 18F(a)], the sum of which is nearly 10 times R_1 . The whole wave front gives $\frac{1}{2}R_1$, so that using only 10 exposed zones, we obtain an amplitude at P which is 20 times as great as when the plate is removed. The intensity is therefore 400 times as great. If the odd zones are covered, the amplitudes R_2, R_4, R_6, \dots will give the same effect. The object and image distances obey the ordinary lens formula, the focal length being the image distance for the object at infinity, namely,

$$b = \frac{s_n^2}{n\lambda} = \frac{s_1^2}{\lambda}$$

There are fainter foci for such a zone plate at distances $b/3, b/5, b/7, \dots$, owing to single zones acting in groups of 3, 5, 7, \dots .

18.6. Vibration Curve for Circular Division of the Wave Front. Our consideration of the vibration curve in the Fraunhofer diffraction by a single slit (Sec. 15.4) was based upon the division of the plane wave front into infinitesimal elements of area which were actually strips of infinitesimal width parallel to the length of the diffracting slit. The vectors representing the contributions to the amplitude from these elements were found to give an arc of a circle. This so-called *strip division* of the wave front is appropriate when the source of light is a narrow slit and the diffracting aperture rectangular. The strip division of a divergent wave

front from such a source will be discussed below (Sec. 18.7). The method of division of the spherical wave front appropriate to the above problem of diffraction by circular apertures and obstacles consists of dividing the wave into infinitesimal circular *zones*.

Let us consider first the amplitude diagram when the first half-period zone is divided into eight subzones, each constructed in a manner similar to that used for the half-period zones themselves. We make these subzones by drawing circles on the wave front (Fig. 18C) which are distant

$$b + \frac{1}{8} \frac{\lambda}{2}, b + \frac{2}{8} \frac{\lambda}{2}, b + \frac{3}{8} \frac{\lambda}{2}, \dots, b + \frac{\lambda}{2}$$

from P . The light arriving at P from various points in the first subzone will not vary in phase by more than $\pi/8$. The resultant of these may be

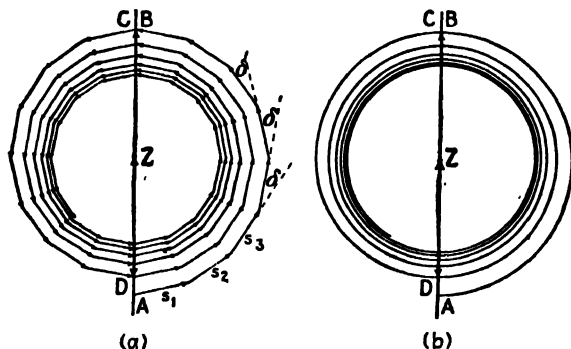


FIG. 18K. Vibration spiral for Fresnel half-period zones of a circular opening.

represented by the vector s_1 in Fig. 18K(a). To this is now added s_2 , the resultant amplitude due to the second subzone, then s_3 due to the third subzone, etc. The magnitudes of these vectors will decrease very slowly as a result of the obliquity factor. The phase difference δ between each successive one will be constant and equal to $\pi/8$. Addition of all eight subzones yields the vector AB as the resultant amplitude from the first half-period zone. Continuing this process of subzoning to the second half-period zone, we find CD as the resultant for this zone, and AD as that for the sum of the first two zones. These vectors correspond to those of Fig. 18F. Succeeding half-period zones give the rest of the figure, as shown.

The transition to the vibration curve of Fig. 18K(b) results from increasing indefinitely the number of subzones in a given half-period zone. The curve is now a *vibration spiral*, eventually approaching Z as an infinite number of turns are included. Any one turn is very nearly

a circle, but does not quite close because of the slow decrease in the magnitudes of the individual amplitudes. The significance of the series of decreasing amplitudes, alternating in sign, used in Sec. 18.2 for the half-period zones, becomes clearer when we keep in mind the curve of Fig. 18K(b). It has the additional advantage of allowing us to determine directly the resultant amplitude due to any fractional number of zones. It should be mentioned in passing that the resultant amplitude AZ , which is just half the amplitude due to the first half-period zone, turns out to be from this treatment, 90° in phase *behind* the light from the center of the zone system. This cannot be true, since it is impossible to alter the resultant phase of a wave merely by the artifice of dividing it into zones and then recombining the effects of these. The discrepancy is a defect of Fresnel's theory resulting from the approximations made therein, and does not occur in the more general mathematical treatment given by Kirchhoff.*

18.7. Cylindrical Wave Fronts. The wave front from a point source is spherical, becoming plane at an infinite distance from the source. In the case of light spreading out from a slit, the envelope of the secondary wavelets is cylindrical, with the slit as the axis of the cylinder. The

Fresnel diffraction phenomena produced when such a wave front passes through an aperture having straight edges parallel to the slit, or by an obstacle having such edges, will now be investigated in some detail.

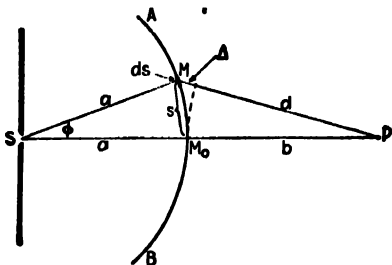


FIG. 18L. Geometry of optical path for the derivation of the Fresnel integrals.

Since the cross section of this system, as shown in the figure, is the same in any plane perpendicular to S , it will be sufficient to consider points S and P and the circular section AB of the wave in the plane of the figure. The secondary wavelets from the various elements ds of the wave front will arrive at P in different phases, having traveled different distances MP . It is required to find the resultant amplitude at P due to the wavelets from the whole wave front. Let the vibration of all points in the wave front AB be represented by

$$y = r \sin \frac{2\pi}{T} t$$

The vibration at P contributed by an element ds at M_0 will be

$$dy = \frac{r ds}{b} \sin 2\pi \left(\frac{t}{T} - \frac{b}{\lambda} \right)$$

taking the amplitude as proportional to ds/b and the phase as retarded by $(2\pi/\lambda)b$. From an element ds at a distance s above M_0 (measured along the wave front) we have

$$\begin{aligned} dy &= \frac{r ds}{d} \sin 2\pi \left(\frac{t}{T} - \frac{d}{\lambda} \right) = \frac{r ds}{d} \sin 2\pi \left(\frac{t}{T} - \frac{b + \Delta}{\lambda} \right) \\ &= \frac{r ds}{d} \sin 2\pi \left[\left(\frac{t}{T} - \frac{b}{\lambda} \right) - \frac{\Delta}{\lambda} \right] \end{aligned}$$

We are here neglecting the obliquity factor, and this will be justified later. Since Δ is different for every point on the wave front, it is convenient to separate this factor using the formula

$$\sin(A - B) = \sin A \cos B - \cos A \sin B \quad (18i)$$

This gives

$$dy = \frac{r}{d} \sin 2\pi \left(\frac{t}{T} - \frac{b}{\lambda} \right) \cos 2\pi \frac{\Delta}{\lambda} ds - \frac{r}{d} \cos 2\pi \left(\frac{t}{T} - \frac{b}{\lambda} \right) \sin 2\pi \frac{\Delta}{\lambda} ds$$

To sum up the contributions from all elements of the wave front between M_0 and M , we integrate this expression between the limits zero and s , assuming that d is essentially a constant:

$$\begin{aligned} y &= \frac{r}{d} \sin 2\pi \left(\frac{t}{T} - \frac{b}{\lambda} \right) \int_0^s \cos \frac{2\pi}{\lambda} \Delta ds \\ &\quad - \frac{r}{d} \cos 2\pi \left(\frac{t}{T} - \frac{b}{\lambda} \right) \int_0^s \sin \frac{2\pi}{\lambda} \Delta ds. \quad (18j) \end{aligned}$$

Now since the sum of any number of cosine functions of the same frequency is another cosine function, and similarly for the sines, we may let

$$\frac{r}{d} \int_0^s \cos \frac{2\pi}{\lambda} \Delta ds = R \cos \theta \quad (18k)$$

and

$$\frac{r}{d} \int_0^s \sin \frac{2\pi}{\lambda} \Delta ds = R \sin \theta \quad (18l)$$

where R and θ remain to be evaluated. Eq. 18j then becomes

$$y = R \sin 2\pi \left(\frac{t}{T} - \frac{b}{\lambda} \right) \cos \theta - R \cos 2\pi \left(\frac{t}{T} - \frac{b}{\lambda} \right) \sin \theta$$

and, again applying Eq. 18i,

$$y = R \sin 2\pi \left(\frac{t}{T} - \frac{b}{\lambda} - \frac{\theta}{2\pi} \right)$$

This represents a vibration having the same period as the wave, but a different amplitude and phase constant. The phase constant is of little interest, but we wish to know the magnitude of the new amplitude R , the square of which gives the intensity. This can be found by squaring Eqs. 18k and 18l and adding. The results

$$R^2(\cos^2 \theta + \sin^2 \theta) = \left[\frac{r}{d} \int_0^s \cos \frac{2\pi}{\lambda} \Delta ds \right]^2 + \left[\frac{r}{d} \int_0^s \sin \frac{2\pi}{\lambda} \Delta ds \right]^2 \quad (18m)$$

Now it remains to find Δ in terms of s . This can be done from the geometry of Fig. 18L. Applying the law of cosines to the triangle SPM , we may write

$$\begin{aligned} d^2 &= (a + b)^2 + a^2 - 2a(a + b) \cos \phi \\ &= 2a^2(1 - \cos \phi) + 2ab(1 - \cos \phi) + b^2 \end{aligned} \quad (18n)$$

Using the relation $1 - \cos \phi = 2 \sin^2 (\phi/2)$, this may be written

$$d^2 = 4a^2 \sin^2 \frac{\phi}{2} + 4ab \sin^2 \frac{\phi}{2} + b^2$$

Further simplification is possible if we assume the angle ϕ small enough so that we may put $\sin \phi = \phi$. The results

$$d^2 = a^2 \phi^2 + ab \phi^2 + b^2$$

This equation contains a relation between Δ and s , since $d = b + \Delta$ and $\phi = s/a$. Squaring this value of d , we find

$$d^2 = b^2 + 2b\Delta + \Delta^2$$

Here we may drop the last term, since Δ is negligible compared to b . The two expressions for d^2 are then equated, and s/a substituted for ϕ , giving

$$b^2 + 2b\Delta = \left(\frac{s}{a} \right)^2 (a^2 + ab) + b^2$$

and finally

$$\Delta = s^2 \left(\frac{a + b}{2ab} \right) \quad (18o)$$

This is the required function of s to be substituted in Eq. 18m. However, it is more convenient to express the intensity in terms of another

variable v , defined by the equation

$$s = v \sqrt{\frac{ab\lambda}{2(a+b)}} \quad (18p)$$

Then, from Eqs. 18o and 18p,

$$\frac{2\pi}{\lambda} \Delta = \frac{2\pi}{\lambda} \frac{a+b}{2ab} r^2 \frac{ab\lambda}{2(a+b)} = \frac{\pi r^2}{2} \quad (18q)$$

and, differentiating Eq. 18p,

$$ds = \sqrt{\frac{ab\lambda}{2(a+b)}} dv$$

Substituting these values in Eq. 18m,

$$R^2 = \frac{r^2}{d^2} \frac{ab\lambda}{2(a+b)} \left\{ \left[\int_0^v \cos \frac{\pi v^2}{2} dv \right]^2 + \left[\int_0^v \sin \frac{\pi v^2}{2} dv \right]^2 \right\} \quad (18r)$$

Hence the intensity becomes

$$I = \text{const.} \times (x^2 + y^2) \quad (18s)$$

where x and y are the integrals,

$$x = \int_0^v \cos \frac{\pi v^2}{2} dv \quad (18t)$$

$$y = \int_0^v \sin \frac{\pi v^2}{2} dv \quad (18u)$$

They are known as *Fresnel's integrals*. The integration of these quantities gives infinite series which may be evaluated in several ways. Although the actual evaluation is too complicated to be given here,* we have included a table of their numerical values (Table 18I). We shall describe below (Sec. 18.14) the use of this table in the computation of diffraction patterns.

18.9. Vibration Curve for Strip Division. Cornu's Spiral. Figure 18M is a curve in which the two Fresnel integrals are plotted against each other, with x as abscissas and y as ordinates. This curve is known as *Cornu's† spiral* and, as we shall show, constitutes the vibration curve for a cylindrical wave front. Let us deduce this vibration curve by a graph-

* For the methods of evaluating Fresnel's integrals, see R. W. Wood, "Physical Optics," 2d ed., p. 247, The Macmillan Company, New York, 1921.

† A. Cornu (1841-1902). Professor of experimental physics at the École Polytechnique, Paris.

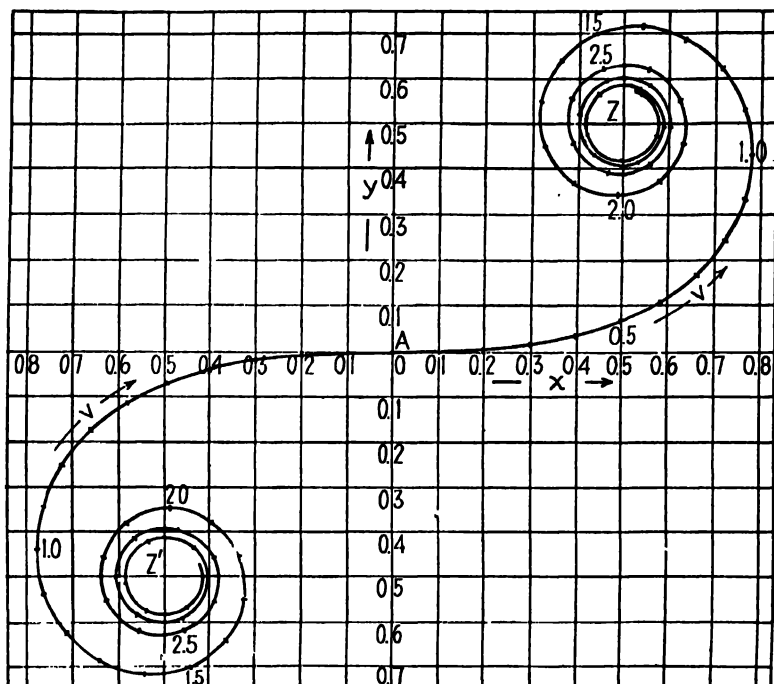


FIG. 18M. Cornu's spiral, a plot of the Fresnel integrals.

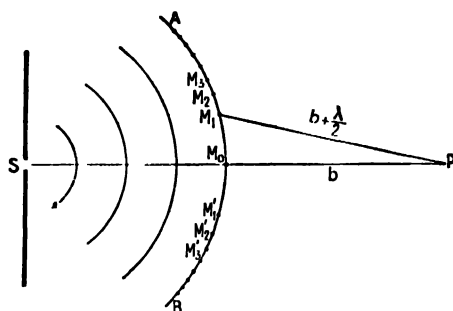


FIG. 18N. Illustrating strip division of cylindrical wave front.

ical method similar to that used in Sec. 18.6 for the circular division of the wave front.

The appropriate method of constructing half-period elements on a cylindrical wave front consists of dividing the wave front into strips, the edges of which are each successively one-half wavelength farther from the point P (Fig. 18N). Thus the points M_0, M_1, M_2, \dots on the circular section of the cylindrical wave are at distances $b, b + (\lambda/2),$

TABLE 18I.—TABLE OF FRESNEL INTEGRALS

v	x	y	v	x	y
0 00	0 0000	0 0000	4.50	0 5261	0.4342
0 10	0 1000	0.0005	4 60	0.5673	0.5162
0 20	0 1999	0 0042	4.70	0 4914	0 5672
0 30	0.2994	0.0141	4.80	0 4338	0 4968
0 40	0 3975	0.0334	4.90	0 5002	0.4350
0.50	0.4923	0 0647	5 00	0.5637	0.4992
0 60	0.5811	0.1105	5.05	0.5450	0.5442
0 70	0 6597	0.1721	5.10	0.4998	0.5624
0.80	0 7230	0.2493	5.15	0.4553	0.5427
0 90	0 7618	0.3398	5 20	0 4389	0.4969
1.00	0 7799	0.4383	5.25	0 4610	0 4536
1.10	0 7638	0.5365	5.30	0.5078	0.4405
1 20	0 7154	0.6234	5.35	0.5190	0 4662
1.30	0 6386	0 6863	5.40	0.5573	0.5110
1 40	0 5431	0.7435	5.45	0 5269	0 5519
1 50	0 4153	0.6975	5.50	0.4784	0.5537
1.60	0 3655	0 6389	5.55	0.4456	0 5181
1.70	0.3238	0 5492	5.60	0.4517	0.4700
1.80	0.3336	0.4508	5.65	0 4926	0.4441
1.90	0.3944	0 3731	5.70	0 5385	0.4595
2 00	0.4882	0.3434	5.75	0.5551	0 5049
2.10	0.5815	0.3743	5 80	0 5298	0.5461
2.20	0.6363	0.4557	5.85	0.4819	0.5513
2 30	0.6266	0 5531	5.90	0 4486	0 5163
2.40	0 5550	0 6197	5.95	0.4566	0.4688
2 50	0.4574	0.6192	6.00	0.4995	0.4470
2 60	0.3890	0 5500	6.05	0 5124	0.4689
2 70	0.3925	0.4529	6 10	0.5495	0 5165
2.80	0 1675	0 3915	6.15	0.5146	0.5496
2.90	0 5626	0.4101	6.20	0.4676	0.5398
3.00	0 6058	0 4963	6.25	0.4493	0 4954
3 10	0 5616	0 5818	6.30	0.4760	0 4555
3 20	0.1664	0.5933	6.35	0.5240	0 4560
3 30	0.4058	0.5192	6.40	0.5496	0 4965
3.40	0 4385	0 4296	6.45	0.5292	0.5398
3 50	0.5326	0.4152	6.50	0.4816	0.5454
3 60	0 5880	0.4923	6.55	0 4520	0.5078
3 70	0.5120	0.5750	6.60	0 4690	0.4631
3.80	0.4481	0.5656	6.65	0.5161	0.4549
3.90	0.4223	0.4752	6.70	0 5467	0.4915
4 00	0.4984	0 4204	6.75	0.5302	0.5362
4.10	0.5738	0.4758	6.80	0.4831	0.5436
4.20	0.5418	0.5633	6.85	0.4539	0.5060
4.30	0.4494	0.5540	6.90	0.4732	0.4624
4.40	0.4383	0.4622	6.95	0.5207	0.4591

$b + (2\lambda/2), \dots$ from P . M_0 is on the straight line SP . The half-period strips now stretch along the wave front, perpendicular to the plane of the figure, and have widths M_0M_1, M_1M_2, \dots . Their appearance when viewed from the point P is shown in Fig. 180. We may call this process *strip division* of the wave front.

In the Fresnel zones obtained by circular division, the areas of the

zones were very nearly equal. In the present case of strip division this is by no means the case. The areas of the *half-period strips* are proportional to their widths, and these decrease rapidly as we go out along the wave front from M_0 .

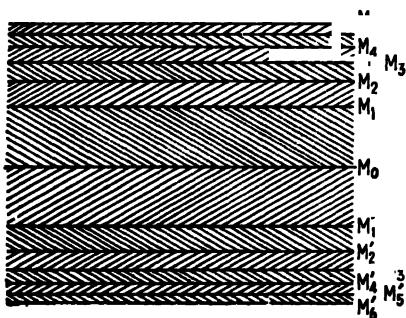


FIG. 180. Fresnel half-period strips for a cylindrical wave front.

zones were very nearly equal. In the present case of strip division this is by no means the case. The areas of the *half-period strips* are proportional to their widths, and these decrease rapidly as we go out along the wave front from M_0 .

The amplitude diagram of Fig. 18P(a) is obtained by dividing the strips into substrips in a manner analogous to that described in Sec. 18.6 for circular zones. Dividing the first half-period strip above M_0 into nine parts, we find that the nine amplitude vectors from the substrips extend from A to B , giving a resultant $R_1 = AB$ for the first half-period strip. The second half-period strip similarly gives those between B and C , with a resultant $R_2 = BC$. Since the amplitudes now decrease rapidly, owing to the decreasing area, R_2 is considerably smaller than R_1 , and their difference in phase is appreciably greater than π . A repetition of this

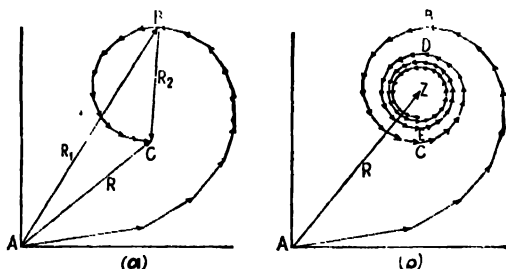


FIG. 18P. Amplitude diagrams for the formation of Cornu's spiral.

process of subdivision for the succeeding strips above M_0 gives the more complete diagram of Fig. 18P(b). Here the vectors are spiraling in toward Z , so that the resultant for all half-period strips in the upper half of the wave becomes AZ .

If now each half-period strip is divided into substrips of infinitesimal width, the diagram of Fig. 18P(b) becomes the upper half of the Cornu

spiral of Fig. 18M. The lower half is obtained in exactly the same way and represents the contributions from all the half-period strips below M_0 . To prove that the vibration curve is identical with Cornu's spiral as plotted from Fresnel's integrals, we note that according to Eq. 18s the intensity is proportional to the sum of the squares of x and y . But x and y are the coordinates of a point on the spiral, and $x^2 + y^2$ is the square of the distance R from the origin to (x, y) . On a vibration curve, the intensity is given by the square of the resultant amplitude, which in the present case (for the upper part of the wave from M_0 to M) is a vector such as AC of Fig. 18P, drawn from the origin to the end of the part of the spiral which is effective. Since this is identical with the distance $R = \sqrt{x^2 + y^2}$, the two methods lead to the same result.

The coordinates of the end points Z and Z' of Cornu's spiral are $(\frac{1}{2}, \frac{1}{2})$ and $(-\frac{1}{2}, -\frac{1}{2})$, respectively. The distance from the origin A to Z is thus $1/\sqrt{2}$, and this represents the amplitude contributed by the upper half of the wave. That contributed by the lower half is $Z'A$ and is also $1/\sqrt{2}$. Squaring the sum of these, we obtain 2 as the intensity due to the whole wave. This numerical value is of no significance—it results from the assumption that the constant in Eq. 18s is unity. It is important, however, to remember that when in the following we calculate relative intensities as R^2 , the square of the resultant amplitude, these are all relative to the value 2 for the unobstructed light. That is, if we wish to express the intensity I as a fraction of the unobstructed intensity I_0 , we have the relation

$$\frac{I}{I_0} = \frac{1}{2} R^2$$

Finally, it is to be noted that distances measured *along the spiral* are proportional to the variable v . This may be seen from the fact that by definition (Eq. 18p) v is proportional to s , the length of the wave that is effective. Since the latter determines the number of infinitesimal amplitudes contributing and hence the length of the arc of the curve, v is proportional to this length. The advantage of using v rather than s as a variable is now apparent, because one scale of v along the curve will apply to all cases. If s were used, there would be a different scale of s for each different set of values of a , b , and λ . The positions of the points $v = 1.0$ and 2.0 on the scale marked in Fig. 18M should be noted. By Eq. 18q they correspond to $\Delta = \lambda/4$ and $\Delta = \lambda$, respectively.

18.10. Straight Edge. In order to appreciate more fully the meaning and use of Cornu's spiral, we shall first consider its application to the simplest problem, namely, that of the diffraction by a straight edge. In Fig. 18Q(a), N represents the section of a screen having a straight edge

parallel to the slit S . In this figure the half-period strips corresponding to the point P on the edge of the geometrical shadow are marked off on the wave front. To find the intensity at P , we note that since the upper half of the wave is effective, the amplitude is a straight line joining A and Z (Fig. 18R) of length $1/\sqrt{2}$. The square of this is $1/2$, so that the intensity at the edge of the shadow is just *one-fourth* of that found above for the unobstructed wave.

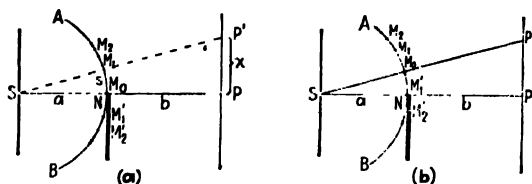


FIG. 18Q. Showing division of a cylindrical wave front AB for Fresnel diffraction at a straight edge N .

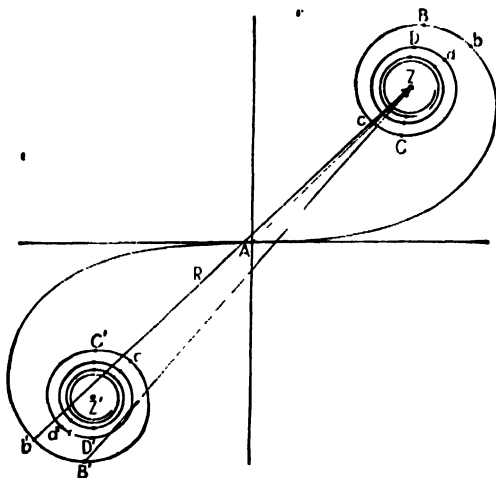


FIG. 18R. Cornu spiral showing resultants for straight-edge diffraction pattern.

Consider next the intensity at the point P' [Fig. 18Q(a)] at a distance x above P . To be specific, let P' lie in the direction SM_1 , where M_1 is the upper edge of the first half-period strip. For this point, the center M_0 of the half-period strips lies on the straight line joining S with P' , and the figure must be reconstructed as in Fig. 18Q(b). The straight edge now lies at the point M'_1 , so that not only all the half-period strips above M_0 are exposed but also the first one below M_0 . The resultant amplitude R is therefore represented on the spiral of Fig. 18R by a straight line joining B' and Z . This amplitude is more than twice that at P , and the intensity R^2 more than four times as great.

Starting with the point of observation P at the edge of the geometrical shadow (Fig. 18Q), where the amplitude is given by AZ , if we move the point steadily upward, the tail of the amplitude vector moves to the left along the spiral, while its head remains fixed at Z . The amplitude will evidently go through a maximum at b' , a minimum at c' , another maximum at d' , etc., approaching finally the value $Z'Z$ for the unobstructed

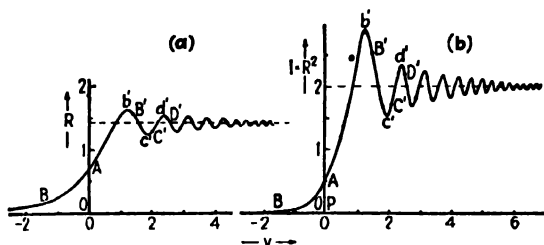


FIG. 18S. Amplitude and intensity contours for Fresnel diffraction at a straight edge.

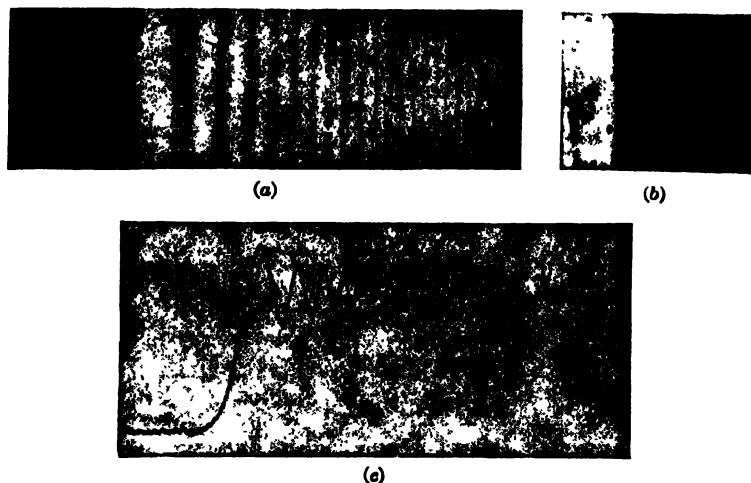


FIG. 18T. Straight-edge diffraction patterns photographed with (a) visible light $\lambda 4300 \text{ \AA}$, and (b) X rays $\lambda 8.33 \text{ \AA}$. (c) Microphotometer trace of (a).

wave. If we go downward from P , into the geometrical shadow, the tail of the vector moves to the right from A , and the amplitude will decrease steadily, approaching zero.

To obtain quantitative values of the intensities from Cornu's spiral, it is only necessary to measure the length R for various values of v . The square of R gives the intensity. Plots of the amplitude and the intensity against v are shown in Figs. 18S(a) and (b) respectively. It will be seen that at the point A , which corresponds to the edge of the geometrical

shadow, the intensity has fallen to one-fourth that for large positive values of v , where it approaches the value for the unobstructed wave. The other letters correspond with points similarly labeled on the spiral, $B', C', D' \dots$, representing the exposure of one, two, three, etc., half-period strips below M_0 . The maxima and minima of the *diffraction fringes* occur a little before these points are reached. For instance, the first maximum at b' is given when the amplitude vector R has the position shown in Fig. 18R. Photographs of the diffraction pattern from a straight edge are shown in Fig. 18T(a) and (b). Pattern (a) was taken with visible light from a mercury arc, and (b) with X rays, $\lambda = 8.33 \text{ \AA}$. Figure 18T(c) is a density trace of the photograph (a), directly above, and was made with a microphotometer.

18.11. Rectilinear Propagation of Light. When we investigate the *scale* of the above pattern for a particular case, the reason for the apparently rectilinear propagation of light becomes clear. Let us suppose that in a particular case $a = b = 100 \text{ cm}$, and $\lambda = 5000 \text{ \AA}$. From Eq. 18p, we then have

$$s = v \sqrt{\frac{ab\lambda}{2(a+b)}} = 0.0354v \text{ cm}$$

This is the distance along the wave front [Fig. 18Q(a)]. To change it to distances l on the screen, we note from the figure that

$$l = \frac{a+b}{a} s = v \sqrt{\frac{b\lambda(a+b)}{2a}} \quad (18v)$$

For the particular case chosen, therefore,

$$l = 2s = 0.0708v \text{ cm}$$

Now in the graph of Fig. 18S(b) the intensity at the point $v = -2$ is only 0.025 or one-eightieth of the intensity if the straight edge were absent. This point has $l = -0.142 \text{ cm}$, and therefore lies only 1.42 mm inside the edge of the geometrical shadow. The part of the screen below this will lie in practically complete darkness, and this must be due to the destructive interference of the secondary wavelets arriving here from the upper part of the wave. In view of the relatively small scale of the diffraction pattern found in this example, our neglect of the obliquity factor and of the variation of amplitude with distance in the derivation of Fresnel's integrals is of small consequence.

18.12. Luminosity of the Diffracting Edge. If the eye is placed at a point such as P' , Fig. 18U, inside the geometrical shadow of a straight edge, one sees a short, bright line of light along the edge N . This means

that the light reaching P' consists of cylindrical wave fronts BB' apparently originating at N . Nothing in the treatment of Sec. 18.10 would indicate the existence of such waves, and the fact that they are observed must mean that they constitute the net effect of the entire exposed wave front extending above the point N . No evidence of their presence was obtained above, because we were there concerned only with amplitudes and intensities and neglected the matter of phases. But the shape of the wave fronts is determined by the variation with the distance below the point P of the resultant phase of the light reaching the screen. The phase for any one point can be found from Cornu's spiral by the angle that the corresponding amplitude vector makes with the x axis. A detailed analysis of its variation with the position of P' , which we shall

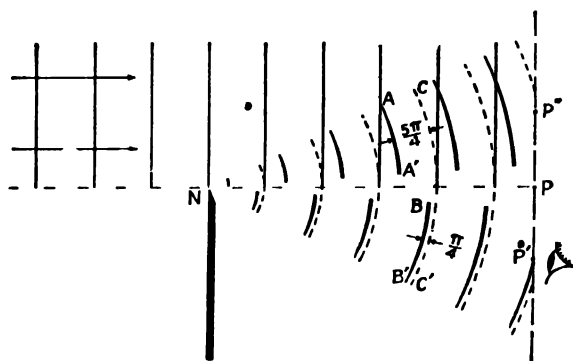


FIG. 18U. Observation of the luminous edge. For the sake of simplicity the parts of the waves near the line NP have been omitted.

not reproduce here,* shows that this variation has the form expected for cylindrical waves the axis of which lies along the straight edge. This result holds only if we are not too close to P and if the wavelength is relatively much smaller than is indicated in Fig. 18U. If an actual slit source were located at N , a representative short section of arc slid along Cornu's spiral will show that the cylindrical waves spreading out from N would have phase relations corresponding to the waves shown by broken lines line CC' . For the straight edge, however, the deflected waves set up by the whole effective wave front have phase relations corresponding to the arcs AA' and BB' . Cornu's spiral shows that below the geometrical shadow P the phase is retarded by $\pi/4$, while above it is retarded by $5\pi/4$. In the region near P the wave fronts, although not shown, are smooth and continuous, connecting such points as A' and B .

Luminosity of the edge of the diffracting obstacle is a general phenomenon, and may be observed, for example, by the light at the center of the shadow of an opaque disk or strip, and even in the region outside the geometrical shadow of a straight edge. If the eye is placed at P'' , Fig. 18U, one observes both the primary source and the luminous edge. The diffraction fringes in this region can be interpreted as due to interference between the light from these two sources. In doing so, one must take into account a retardation of phase of $5\pi/4$ by which the deflected waves lag behind the primary waves. In view of the fact that the diffracted light always appears to come from the edge itself, it is understandable that the early investigators of diffraction first thought of interactions between the light and the material of the edge itself (Sec. 18.1). Young, for example, tried to explain the straight-edge fringes as due to interference between the direct light and that reflected from the

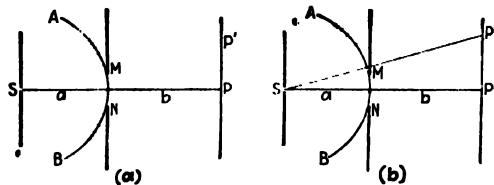


FIG. 18V. Showing division of a cylindrical wave front AB for Fresnel diffraction by a single slit.

edge. It was not until Fresnel's work that the true significance of the luminous edge was appreciated.

18.13. Single Slit. We next consider the Fresnel diffraction of a single slit with sides parallel to a narrow source slit S [Fig. 18V(a)]. By the use of Cornu's spiral we wish to determine the distribution of the light on the screen PP' . With the slit located as shown, each side acts like a straight edge to screen off the outer ends of the wave front AB . We have already seen in the last section how to investigate the pattern from a single straight edge, and the method used there is readily extended to the present case. With the slit in the central position of Fig. 18V(a), the only light arriving at P is that due to the wave front in the interval $\Delta s = MN$. In terms of Cornu's spiral we must now determine what length Δv corresponds to the slit width Δs . This is done by Eq. 18p, using Δv for v and Δs for s . Let $a = 100$ cm, $b = 400$ cm, $\lambda = 4000 \text{ \AA} = 0.00004$ cm, and slit width $\Delta s = 0.02$ cm. Substituting in Eq. 18p, we obtain $\Delta v = 0.5$. The resultant amplitude at P is then given by a chord of the spiral, the arc of which has a length $\Delta v = 0.5$. Since the point of observation P is centrally located, this arc

will start at $v = -0.25$ and run to $v = +0.25$. This resultant $R \cong 0.5$ when squared gives the intensity at P .

If we now wish the intensity at P' [Fig. 18V(b)], the picture must be revised by redividing the wave front as shown. With the point of observation at P' , the same length of wave front, $\Delta s = 0.02$ cm, is exposed, and therefore the same length of the spiral, $\Delta v = 0.5$, is effective. This section on the lower half of the wave front will, however, correspond to a new position of the arc on the lower half of the spiral. Suppose that it is represented by the arc jk in Fig. 18W. The resultant

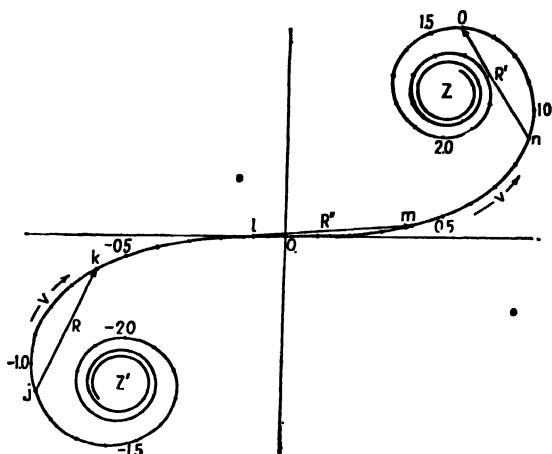


FIG. 18W. Cornu's spiral showing equal arc lengths v .

amplitude is proportional to the chord R , and the square of this gives the relative intensity. Thus to get the variation of intensity along the screen of Fig. 18V, we slide a piece of the spiral of *constant* length $\Delta v = 0.5$ to various positions and measure the lengths of the corresponding chords to obtain the amplitudes. In working a specific problem, the student may make a straight scale marked off in units of v to tenths, and measure these distances on an accurate plot such as Fig. 18M, using the scale of v on the spiral to obtain the constant length Δv of the arc. The results should then be tabulated in three columns, giving v , R , and R^2 . In the first column, the value of v for the central point of the arc whose chord R is being measured should be tabulated. For example, if the interval from $v = 0.9$ to $v = 1.4$ is measured (Fig. 18W), the average value $v = 1.15$ is tabulated against $R = 0.43$.

Photographs of a number of Fresnel diffraction patterns for single slits of different widths are shown in Fig. 18X with the corresponding

intensity curves beside them. These curves have been plotted by the use of Cornu's spiral. It is of interest to note in each curve the indicated positions of the edges of the geometrical shadow of the slit. Very little light falls outside these points. For a very narrow slit like the first of these where $\Delta v = 1.5$, the pattern greatly resembles the Fraunhofer diffraction pattern for a single slit. The essential difference between the two (cf. Fig. 15D) is that here the minima do not come quite to zero except at infinitely large v . The small single-slit pattern at the top was taken with X rays of wavelength 8.33 Å, while the rest were

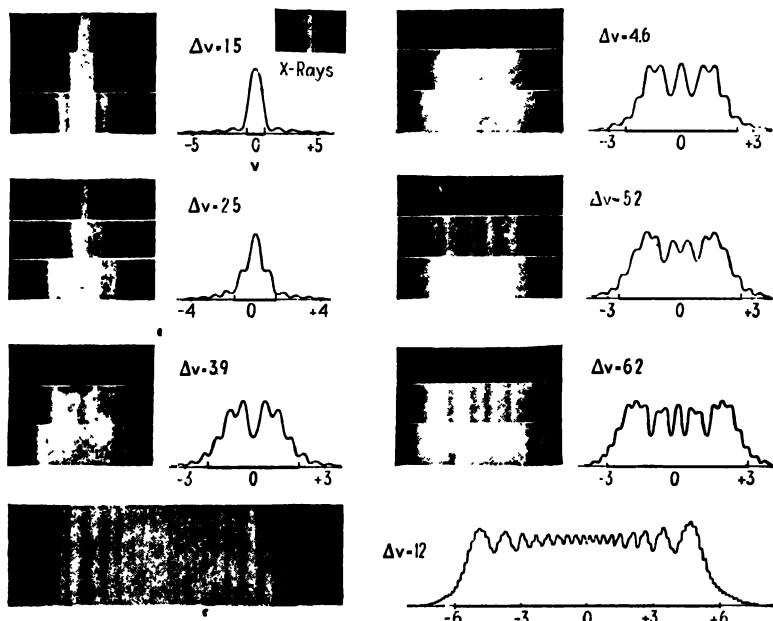


FIG. 18X. Fresnel diffraction of visible light by narrow slits. (X-ray pattern after Kellstrom.)

taken with visible light of wavelength 4358 Å. As the slit becomes wider, the fringes go through very rapid changes, approaching for a wide slit the general appearance of two opposed straight-edge diffraction patterns. The small closely spaced fringes superimposed on the main fringes at the outer edges of the last figure are clearly seen in the original photograph and may be detected in the reproduction.

18.14. Use of Fresnel's Integrals in Solving Diffraction Problems. The tabulated values of Fresnel's integrals in Table 18I may be used for specific problems in place of the plotted spiral. Although it is more tedious, this method is the most accurate. For an interval $\Delta v = 0.5$, for

example, the two values of x at the ends of this interval are read off and subtracted algebraically to give Δx . The corresponding two values of y are also subtracted to give Δy . These according to Eq. 18s are squared and added to obtain directly the intensity R^2 for the mid-point v . In the case of the straight edge, and others where the number of zones on one end of the interval is not limited, the values of both x and y will be $\frac{1}{2}$ at this end. This is also the case in the next example we shall discuss.

18.15. Diffraction by an Opaque Strip. The shadow cast by a narrow object with parallel sides, such as a wire, may be studied by the use of

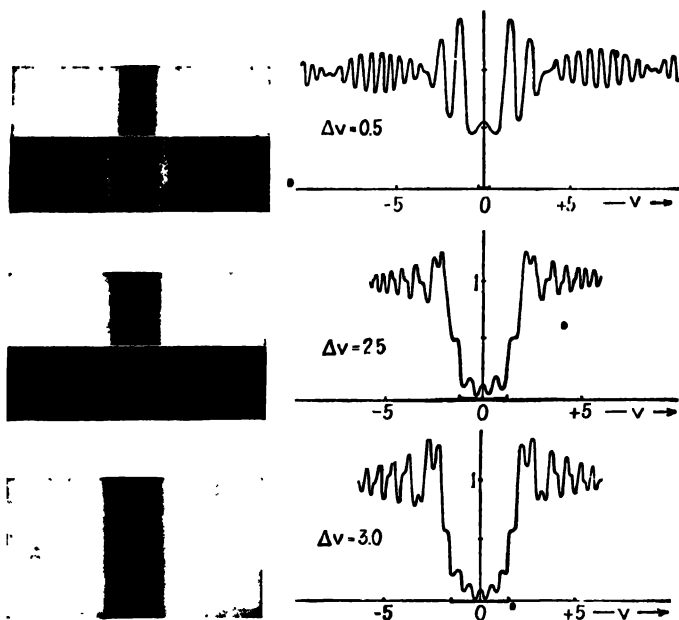


FIG. 18T. Fresnel diffraction by narrow opaque strips.

Cornu's spiral. In the case of a single slit, treated in Sec. 18.13, it was shown how the resultant diffraction pattern is obtained by sliding a fixed length of the spiral, $\Delta v = \text{constant}$, along the spiral and measuring the chord between the two end points. The rest of the spiral out to infinity, *i.e.*, out to Z or Z' on each side of the element in question, was absent owing to the screening by the two sides of the slit. If now the opening of the slit in Fig. 18V(a) is replaced by an object of the same size, and the slit jaws taken away, we have two segments of the spiral to consider. Suppose the obstacle is of such a size that it covers an integral $\Delta v = 0.5$ on the spiral (Fig. 18W). For the position jk the

light arriving at the screen will be due to the two parts of the spiral, one from Z' to j and the other from k to Z . The resultant amplitude due to these two sections is obtained by adding their respective amplitudes as vectors. The lower section gives an amplitude represented by a straight line from Z' to j , with the arrowhead at j . The amplitude for the upper section is represented by a straight line from k to Z with the arrowhead at Z . The vector sum of these two gives the resultant amplitude R , and R^2 gives the intensity for a point v halfway between j and k . Photographs of three diffraction patterns produced by small wires are shown in Fig. 18Y, accompanied by the corresponding curves determined directly from Cornu's spiral.

18.16. Double Slit. As in the case of the single slit, the Fresnel diffraction pattern for a double slit shows marked differences from the Fraunhofer pattern. To study this case we again make use of the Cornu's spiral of Fig. 18W. The treatment of the single slit in Sec. 18.13 and of the opaque strip in Sec. 18.15 lends itself to an easy extension. Knowing the distances a and b as in Fig. 18V, the wavelength λ , the two equal slit widths Δs , and the separation $\Delta s'$ between the centers of the slits, one first calculates the corresponding Δv intervals by means of Eq. 18p. For example, let $a = 100$ cm, $b = 400$ cm, $\lambda = 4000$ Å, $\Delta s = 0.02$ cm, and $\Delta s' = 0.04$ cm. By Eq. 18p each slit is represented on the spiral by two equal arcs of length $\Delta v = 0.5$, the centers of which are separated by an interval of $\Delta v' = 1.0$. Let the two arcs be represented by jk and lm separated by an equal gap kl (Fig. 18W). The resultant amplitude for each of these two arcs, given by the arrows R and R'' , must now be added vectorially to give a total resultant R . The variation in intensity on the screen is again obtained by sliding the arcs, fixed in length and separation, around the spiral, reading off for different points the resultant amplitudes and the corresponding value of v . The positions lm and no , separated by the missing interval mn , illustrate one other location of the arcs. The value of v used for this second position of the three equal sections of the spiral is the central point $v = 0.65$. After tabulation, the values of v may be transformed into distances x on the screen by means of Eq. 18v. Care must be taken when adding the two vectors for each position that the arrows are taken in the right direction. This is determined by the fact that the infinitesimal amplitude vectors of which the spiral is composed start at Z' and end at Z . Hence the arrowhead of any vector must be at the end nearer on the spiral to Z .

The complementary case to the double slit is that of two opaque strips or wires parallel to each other. On the spiral there will now be two absent sections like jk and lm , and three open sections $Z'j$, kl , and mZ .

Three vectors, obtained by joining these three latter pairs of points, must then be added vectorially to give the resultant amplitude R .

18.17. Babinet's Principle. For the general case of any two complementary diffracting screens such as S_1 and S_2 of Fig. 18Z, this principle states that at any place on the observing screen where the intensity in the absence of both S_1 and S_2 is zero, the diffraction pattern will be exactly the same when either S_1 or S_2 is used by itself. In Fresnel diffraction, *i.e.*, in the absence of the lens L , the principle applies only to regions well outside M and N . If, however, L is used to focus an image of the source S at P_0 , this image will be very small in the absence of any diffracting screens, since it will be just the Fraunhofer pattern corresponding to the aperture of the lens. Then Babinet's principle

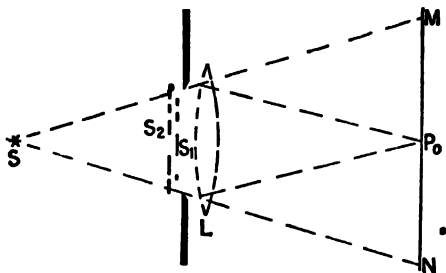


FIG. 18Z. Babinet's principle.

will apply to the whole field MN , except in the immediate neighborhood of P_0 . The principle is proved as follows: Let R_1 be the amplitude produced at any observing point whatsoever by S_1 alone, and R_2 that produced at the same point by S_2 alone. The amplitude existing at this point when both screens are absent must be the vector sum of R_1 and R_2 , because the parts of the wave passed by the two screens are complementary, and these parts taken together constitute the whole wave. Then at any point where $R = 0$, we must have $R_2 = -R_1$. The amplitudes are equal but the phases are opposite. When the amplitudes are squared to obtain intensities, the equality of the latter is proved.

Problems

1. If a large number of opaque disks of equal size are randomly distributed (*e.g.*, by dusting lycopodium powder on a glass plate), and if a distant light source is then viewed through the array, a diffraction disk surrounded by a set of rings is seen, the appearance being very similar to that of the Fraunhofer pattern of a circular aperture (Fig. 15K). (a) By the use of Babinet's principle, explain the existence of such a pattern. (b) If the angular diameter of the first bright ring is 2° for sodium light, what is the diameter of the disks?

2. Monochromatic light of wavelength 5633 Å originates at a distant point source and passes through a circular opening of a diameter which is continuously variable (see Fig. 18*G*). (a) Tabulate the values of the diameter of the hole in centimeters at which the maxima and minima occur at a point on the axis of the hole and 2 m behind it. (b) For a hole 3 mm in diameter, tabulate the distances in centimeters from the hole, along the axis, at which the maxima and minima occur.

3. Compute the radius of a zone plate having a focal length of 2.5 m for white light ($\lambda = 5550$ Å). Assume the plate to have 16 open zones, the central one being opaque.

4. A zone plate is drawn and then is copied on a reduced scale so that the diameter of the central zone is 2.3 mm. If a point source of red light, $\lambda 6000$, is placed 600 cm from the zone plate, find the positions of the primary image and of the weaker secondary images.

5. Plot Cornu's spiral from the values of v , x , and y in Table 18I. Use a fairly large sheet of graph paper, so that both halves of the spiral may be included, and mark off the scale of v on the spiral.

6. Using Cornu's spiral from Prob. 5, plot a straight-edge diffraction pattern for which $a = 90$ cm, $b = 300$ cm, and $\lambda = 4350$ Å. Plot intensity I^2 against distance x measured on the screen. What are the values of x for the first three maxima?

7. Using Cornu's spiral from Prob. 5, plot a single-slit diffraction pattern for one of the following slit widths: $\Delta s = 0.2$ mm, 0.4 mm, 0.6 mm, 0.8 mm, 1.0 mm, 1.4 mm, 1.8 mm, 2.4 mm. Assume $a = 200$ cm, $b = 300$ cm, and $\lambda = 1000$ Å.

8. Using Cornu's spiral from Prob. 5, plot the diffraction pattern of an obstacle the same size as one of the slits of Prob. 7.

9. Using Cornu's spiral from Prob. 5, plot the diffraction pattern of a double slit. Assume the intervals on the spiral corresponding to the width of each slit to be $\Delta v = 0.5$, and that corresponding to the opaque strip between them to be $\Delta v = 0.6$.

10. Calculate accurately, by the use of the table of Fresnel's integrals (Table 18I), the intensity in the shadow of a straight edge at a point corresponding to $v = -3.9$. Express the result relative to the unobstructed intensity. Find the distance on the screen of this point from the edge of the geometrical shadow, if $a = 80$ cm, $b = 250$ cm, and $\lambda = 4500$ Å.

11. Determine the intensity at the minimum of the first dark fringe in the straight-edge diffraction pattern, expressing it relative to that of the unobstructed wave. The approximate value of v may first be determined from Cornu's spiral, and then the exact intensities for several values of v in this neighborhood may be computed from the table of Fresnel's integrals. A plot will then give the minimum.

12. Find the maximum intensity obtainable at the center of a single-slit diffraction pattern, as the slit width is varied. To what slit width does this maximum correspond if $a = 100$ cm, $b = 300$ cm, and $\lambda = 4000$ Å? The method suggested in Prob. 11 may also be applied here.

13. By what percentage does the area of the tenth Fresnel half-period zone differ from that of the first, when $b = 2$ m and $\lambda = 6000$ Å?

14. A source slit is mounted on an optical bench, 150 cm from a holder for diffracting screens, and is illuminated with sodium light. Observations are to be made 150 cm behind the screens. Using the values of Fresnel's integrals in Table 18I, calculate the exact intensity, relative to that of the unobstructed beam, (a) at a point 3 mm outside the edge of the geometrical shadow of a straight edge, (b) at the center of the pattern due to a single slit 2 mm wide, and (c) on the edge of the geometrical shadow of an opaque strip 1 mm wide.

15. Let I represent the relative intensity at any point in a single-slit pattern, and let Δx and Δy represent the components of the corresponding amplitude vector. By consideration of the relevant vectors on Cornu's spiral, prove that the relative intensity at the same point in the pattern due to the complementary screen (*i.e.*, of an opaque strip of the same width as the slit) is given by $1 + I - \Delta x - \Delta y$.

16. Investigate the fringes within the shadow of an opaque strip from the standpoint of their being caused by interference of the light from the two luminous edges of the strip acting as two narrow line sources. Take the example $\Delta v = 3.0$ given in Fig. 18I', and see whether the spacing of the fringes corresponds to this hypothesis.

CHAPTER 19

THE VELOCITY OF LIGHT

In the preceding chapters we have found that the interference and diffraction of light can be successfully explained by assuming that light consists of waves. We now turn to another fundamental property of light waves, their velocity of propagation. It is to be expected that waves having a definite frequency will travel with finite and constant velocity in a given medium. Light waves, or in general, electromagnetic waves, are unique in their ability to move through empty space and here the velocity is the same for all frequencies. Hence the velocity of

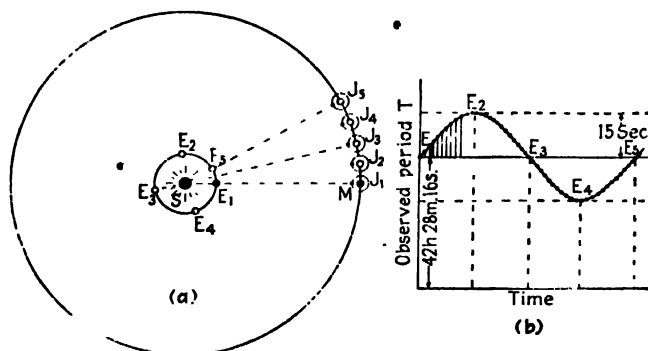


FIG. 19A. Illustrating Römer's astronomical observations on Jupiter's moons, from which the velocity of light was first measured.

light in vacuum, c , is an important constant of nature. Our first object will be to describe the various ways in which this quantity has been accurately measured.

19.1. Römer's Method. Because of the very great velocity of light, it is natural that the first successful measurement was an astronomical one, because here very large distances are involved. In 1676 Römer* studied the times of the eclipses of the satellites of the planet Jupiter. Figure 19A(a) shows the orbits of the earth and of Jupiter around the sun S and that of one of the satellites M around Jupiter. The inner satellite has an average period of revolution $T_0 = 42$ hr 28 min 16 sec, as determined from the average time between two passages into the shadow of the planet. Actually Römer measured the times of *emergence*

* Olaf Römer (1644–1710). Danish astronomer. His work on Jupiter's satellites was done in Paris, and later he was made astronomer royal of Denmark.

from the shadow, while the times of transit of the small black spot representing the shadow of the satellite on Jupiter's surface across the median line of the disk can be still more accurately measured.

A long series of observations on the eclipses of the first satellite permitted an accurate evaluation of the average period T_0 . Römer found that if an eclipse was observed when the earth was at such a position as E_1 [Fig. 19A(a)] with respect to Jupiter J_1 , and the time of a later eclipse was predicted by using the average period, it did not in general occur at exactly the predicted time. Specifically, if the predicted eclipse was to occur about 3 months later, when the earth and Jupiter were at E_2 and J_2 , he found a delay of somewhat more than 10 min. To explain this, he assumed that light travels with a finite velocity from Jupiter to the earth, and that since the earth at E_2 is farther away from Jupiter, the observed delay represents the time required for light to travel the additional distance. His measurements gave 11 min as the time for light to go a distance equal to the radius of the earth's orbit. We now know that 8 min 18 sec is a more nearly correct figure, and combining this with the average distance to the sun 93×10^6 miles, we find a velocity of about 187,000 mi/sec.

It is instructive to inquire how the apparent period of the satellite, *i.e.*, the time between two successive eclipses, is expected to vary throughout a year. If this time could be observed with sufficient accuracy, one would obtain the curve of Fig. 19A(b). We may regard the successive eclipses as light signals sent out at regular time-intervals of 42 hr 28 min 16 sec from Jupiter. Now at all points in its orbit except E_1 and E_2 the earth is changing its distance from Jupiter more or less rapidly. If the distance is increasing, as at E_2 , any one signal travels a greater distance than the preceding one and the observed time between them will be increased. Similarly at E_1 it will be decreased. The maximum variation from the average period, about 15 sec, is the time for light to cover the distance moved by the earth between two eclipses, which amounts to 2.8×10^6 miles. At any given position, the total time delay of the eclipse, as observed by Römer, will be obtained by adding the amounts $T - T_0$ [Fig. 19A(b)], by which each apparent period is longer than the average. For instance, the delay of an eclipse at E_2 , predicted from one at E_1 using the average period, will be the sum of $T - T_0$ for all eclipses between E_1 and E_2 .

19.2. Bradley's* Method. The Aberration of Light. Römer's interpretation of the variations in the times of eclipses of Jupiter's satellites

* James Bradley (1693-1762). At the time professor of astronomy at Oxford. He got his ideas about aberration by a chance observation of the changes in the apparent direction of the wind while sailing on the Thames.

was not accepted until an entirely independent determination of the velocity of light was made by the English astronomer Bradley in 1727. Bradley discovered an apparent motion of the stars which he explained as due to the motion of the earth in its orbit. This effect, known as *aberration*, is quite distinct from the well-known displacements of the nearer stars known as parallax. Because of parallax, these stars appear to shift slightly relative to the background of distant stars when they are viewed from different points in the earth's orbit, and from these shifts the distances of the stars are computed. Since the apparent displacement of the star is opposite to that of the *position* of the earth, the effect of parallax is to cause the star which is observed in a direction perpendicular to the plane of the earth's orbit to move in a small circle

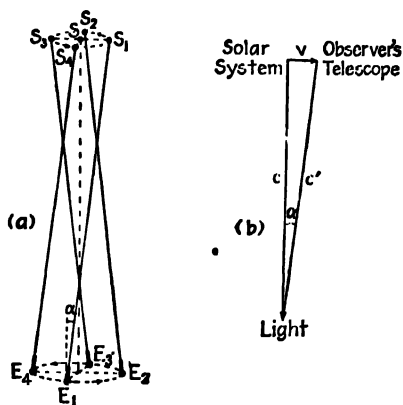


FIG. 19B. Illustrating Bradley's astronomical observations of the aberration of light.

Bradley's explanation of this effect was that the apparent direction of the light reaching the earth from a star is altered by the motion of the earth in its orbit. The observer and his telescope are being carried along with the earth at a velocity of about 18.5 mi/sec, and if this motion is perpendicular to the direction of the star, the telescope must be tilted slightly toward the direction of motion from the position it would have if the earth were at rest. The reason for this is much the same as that involved when a person walking in the rain must tilt his umbrella forward to keep the rain off his feet. In Fig. 19B(b), let the vector v represent the velocity of the telescope relative to a system of coordinates fixed in the solar system, and c that of the light relative to the solar system. We have represented these motions as perpendicular to each other, as would be the case if the star lay in the direction shown in Fig. 19B(a). Then the velocity of the light relative to the earth has the

with a phase differing by π from the earth's motion. The angular diameters of these circles are very small, being not much over 1 second of arc for the nearest stars. Aberration also causes the stars observed in this direction apparently to move in circles, but here the circles have an angular diameter of about .41 seconds, and they are the same for all stars, whether near or distant. Furthermore, the displacements are always in the direction of the earth's velocity, so that the circular motions are $\pi/2$ different in phase from the earth's motion [Fig. 19B(a)].

direction of c' , which is the vector difference between c and v . This is the direction in which the telescope must be pointed to observe the star image on the axis of the instrument. We thus see that when the earth is at E_1 the star S has the apparent position S_1 , when it is at E_2 , the apparent position is S_2 , etc. If S were not in a direction perpendicular to the plane of the earth's orbit, the apparent motion would be an ellipse rather than a circle, but the major axis of the ellipse would be equal to the diameter of the circle in the above case.

It will be seen from the figure that the angle α , which is the angular radius of the apparent circular motion, is given by

$$\tan \alpha = \frac{v}{c} \quad (19a)$$

Recent measurements of this so-called *angle of aberration* give a mean value $\alpha = 20.479'' \pm 0.008$ as the angular radius of the apparent circular

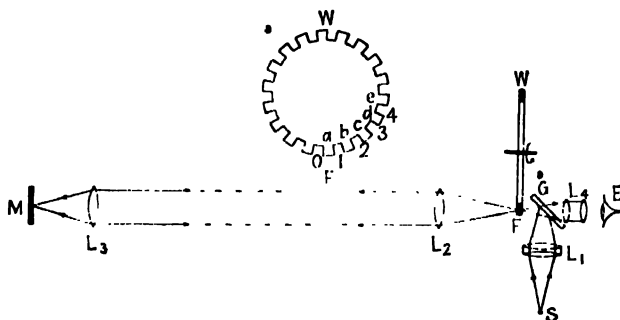


FIG. 19C. Fizeau's experimental arrangement used in the first terrestrial determination of the velocity of light.

orbit. Combining this with the known velocity v of the earth in its orbit, we obtain $c = 186,233$ mi/sec, or 299,714 km/sec. This value agrees to within its experimental error with the more accurate results obtained by the latest measurements of the velocity of light by direct methods, the principles of which we shall now describe.

19.3. Fizeau's Terrestrial Method. Fizeau,* in 1849, first succeeded in measuring c by a method not involving astronomical observations, *i.e.*, one in which the light path was on the earth's surface. The principle of his determination was the obvious one of sending out a brief flash of light and measuring the time for this to travel to a distant mirror and back to the observer. This was accomplished with the apparatus shown in Fig. 19C. The cogwheel WF is rotated at high speed so that

* H. L. Fizeau (1819-1896). Born of a wealthy French family, he was financially independent to pursue his hobby—the velocity of light. His experiments were carried out in Paris, the light traveling between Montmartre and Suresnes.

it cuts the light beam passing through the rim at F into a series of short flashes. A flash is sent out each time the wheel is in such a position that the light can pass between two cogs. It is then rendered parallel by the lens L_2 and focused by L_3 on a plane mirror M . In Fizeau's experiments the distance MF was 5.36 miles. After reflection from M , the flash of light retraces its path, and is again focused by L_2 on the rim of the wheel. If during the time that the light has traveled from F to M and back the wheel has turned to such a position that a cog is interposed at F , this flash will be cut out, and the same will be true of any other flash.

With the wheel at rest in such a position that the light traverses the opening O between two cogs (Fig. 19C', center), the observer at E will see the image of the light source at F by means of the eyepiece L_4 , focused on F through the half-silvered mirror G . If the wheel is now rotated with increasing speed, a state will be reached in which the light passing O is stopped by a , that passing 1 is stopped by cog b , etc., and the image will be completely eclipsed. A further increase in speed will cause the light to reappear when these flashes pass through openings $1, 2, \dots$, and a second eclipse will occur where they are stopped by b, c, \dots . Fizeau's wheel had 720 cogs, and since the light path was 2×5.36 or 10.72 miles, the wheel had to turn through $\frac{1}{720}$ of a revolution in $10.72/c$ sec to produce the first eclipse. Hence the first eclipse should occur at a speed of $c/(10.72 \times 1440)$ rev/sec, and the others at 3, 5, 7, \dots times this speed. Fizeau observed the first eclipse at 12.6 rev/sec, giving $c = 194,600$ mi/sec or 313,300 km/sec.

That this is appreciably higher than the values obtained by the astronomical methods is not surprising, in view of the difficulties of the experiment. With Fizeau's arrangement, the determination of the exact condition of total eclipse caused the principal uncertainty. The experimental conditions were later improved by Cornu, and by Young and Forbes. The latter overcame the above difficulty by placing another lens and mirror, identical with L_3 and M , at a somewhat greater distance. The two images thus formed were observed simultaneously, and instead of measuring the conditions of eclipse or of maximum in either image they measured the speed of the cogwheel at the time the two images appeared to be of equal intensity. The eye is very sensitive to the detection of slight *differences* in intensity of adjacent images, so this measurement could be made more accurately. Their result* was 301,400 km/sec.

19.4. Rotating-mirror Method. This is a second terrestrial method, originally suggested by Arago* and first applied successfully by Fizeau and Foucault† independently in 1850. The principle of these early determinations is illustrated in Fig. 19D. Light from the source S traverses the plane glass plate G and, after reflection from the plane mirror R , is focused by the lens L on a stationary concave mirror M . If R is also stationary, the light retraces its path and an image of S is formed at E by partial reflection in G .

If now R is rotated at high speed about an axis perpendicular to the plane of the figure, it will have turned through a small angle α by the time the light has returned from M . The reflected beam will then be turned through 2α , and a displaced image E' will be produced by L . The displacement EE' obviously depends on the angular velocity of R

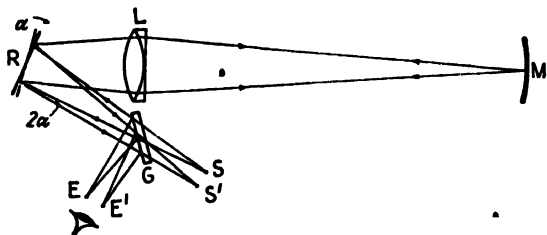


FIG. 19D. Rotating-mirror apparatus used by Foucault in measuring the velocity of light. and on the distances RM and RGE , and if these quantities are known the velocity of light may be found.

In the final measurements of Foucault, RM was 20 m and essentially equal to the radius of curvature LM of the mirror M . The displacement EE' was only 0.7 mm but could be measured by the micrometer eyepiece to within 0.005 mm. Foucault's result for the velocity of light was roughly 298,000 km/sec. The accuracy of the determination by the rotating-mirror method was later greatly improved in the experiments of Cornu, Newcomb,‡ and Michelson. The chief improvement in the later work lay in the use of a greater light path. This was limited in Foucault's arrangement by the loss of intensity in the image when the

* D. F. J. Arago (1786–1853). Noted Parisian astronomer and physicist. He is principally known for his work on the interference of polarized light (Chap. 26) and on electromagnetism in conjunction with Ampère.

† J. L. Foucault (1819–1868). Between 1845 and 1849 Foucault collaborated with Fizeau, but owing to difference of opinion they afterward worked separately. Foucault is also known for his demonstration of the rotation of the earth by a pendulum and for the Foucault knife-edge test. His researches on the velocity of light in water (Sec. 19.10) constituted his thesis for the doctorate, presented in 1851.

‡ Simon Newcomb (1835–1909). Distinguished American astronomer, associated with the U.S. Naval Observatory and the Johns Hopkins University.

distance RM was made large. The rotating beam from R is returned by M only during the small fraction of the time that it is sweeping across M . This difficulty was overcome in Michelson's work by using a lens L of larger focal length, and increasing the distance RL until R and M were nearly conjugate foci of L . With S fairly close to R , and a lens L of sufficiently long focus, the mirror M could now be placed several miles away. Another improvement adopted by Newcomb and Michelson was the replacement of the plane mirror R by one having four or more reflecting faces (Fig. 19E). This also resulted in a gain of intensity in the image.*

19.5. Michelson's Later Experiments. We shall not describe the successive experiments in which the determination of c by rotating mirrors was steadily improved. The numerical results with their estimated

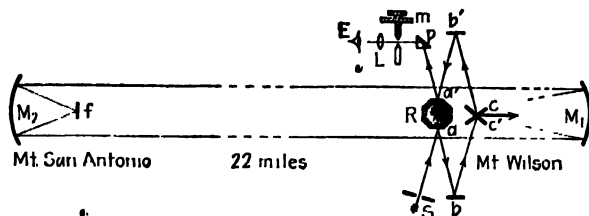


FIG. 19E. Michelson's arrangement used for determining the velocity of light (1926).

errors will be given later in Table 19I. Instead, we shall consider two of the latest determinations by this method, both of which were instituted by Michelson. The first of these is a series of measurements carried out at the Mt. Wilson Observatory in 1926, which constituted a considerable improvement in the accuracy of this method.

The form of the apparatus finally adopted is shown in Fig. 19E. Light from a Sperry arc S passes through a narrow slit and is reflected from one face of the octagonal rotating mirror R . Thence it is reflected from the small fixed mirrors b and c to the large concave mirror M_1 (30-ft focus, 2-ft aperture). This gives a parallel beam of light, which travels 22 miles from the observing station on Mt. Wilson to a mirror M_2 , similar to M_1 , on the summit of Mt. San Antonio. M_2 focuses the light on a small plane mirror f , whence it returns to M_1 and, by reflection from c' , b' , a' , and p , to the observing eyepiece L .

Various rotating mirrors, having 8, 12, and 16 sides, were used, and in each case the mirror was driven by an air blast at such a speed that during the time of transit to M_2 and back (0.00023 sec) the mirror turned through such an angle that the next face was presented at a' . For an

* For further discussion of these methods, see Preston, *op. cit.*, pp. 543-552.

octagonal mirror, the required speed of rotation was about 528 rev/sec. The speed was adjusted by a small counterblast of air until the image of the slit was in the same position as when R was at rest. The exact speed of rotation was then found by a stroboscopic comparison with a standard electrically driven tuning fork, which in turn was calibrated with an invar pendulum furnished by the U.S. Coast and Geodetic Survey. This Survey also measured the distance between the mirrors M_1 and M_2 with remarkable accuracy by triangulation from a 40-km base line, the length of which was determined to an estimated error of 1 part in 11 million, or about $\frac{1}{8}$ in.*

The results of the measurements published in 1926 comprised eight values of the velocity of light, each the average of some 200 individual determinations with a given rotating mirror. These varied between the extreme values of 299,756 and 299,803 km/sec and yielded the average value of $299,796 \pm 4$ km/sec. Michelson also made some later measurements with the distant mirror, on the summit of a mountain 82 miles away, but because of bad atmospheric conditions, these were not considered reliable enough for publication.

19.6. Correction to Vacuum. In the preceding discussion we have assumed that the measured velocity in air is equal to that in a vacuum. That is not exactly true, since the index of refraction $n = c/v$ is slightly greater than unity. With white light the effective value of n for air under the conditions existing in Michelson's experiments was 1.000225. Hence the velocity in vacuum $c = nv$ was 67 km/sec greater than v , the measured value in air. This correction has been applied in the final results quoted above. A difficulty which becomes important where measurements as accurate as those of Michelson are concerned is the uncertainty of the exact conditions of temperature and pressure of the air in the light path. Since n depends on these conditions, the value of the correction to vacuum also becomes somewhat uncertain. The latest measurements of the velocity of light, which we shall now describe, were designed to eliminate this uncertainty.

19.7. The 1-mile Evacuated Pipe Experiment. This experiment was begun in 1929 by Michelson, Pease, and Pearson, but was not completed until after Michelson's death in 1931. The object was to measure the velocity of light in a long evacuated pipe, using a method similar to that of Michelson's measurements described above. An iron pipe 3 ft in diameter and 1 mile long was constructed, with the joints carefully sealed so that a vacuum of from 5.5 to 0.5 mm Hg could be maintained by a single vacuum pump.

The optical arrangement is illustrated in Fig. 19F. Light from the large carbon arc S is focused by L_1 on the slit S_1 , whence it is reflected from one of the 32 faces of the rotating mirror R into the vacuum pipe through the window W . After reflection from the plane mirror Q , it is rendered parallel by the large concave mirror N , of diameter 40 in. and focus 50 ft. This parallel beam travels to the other end of the pipe, where it is reflected at a from the plane mirror M_1 , 2 ft in diameter. Returning to a similar mirror M_2 , it is again reflected at b , and successively at c , d , and e . The mirrors M_1 and M_2 are slightly inclined to each other, so that the light strikes at e perpendicular to the surface of M_1 and retraces its path to N . Then, after reflection from Q , it strikes the

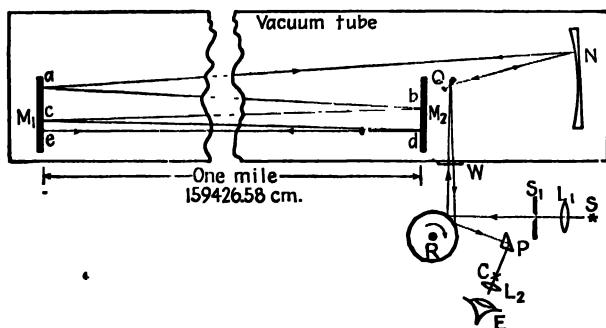


FIG. 19F. Apparatus used by Michelson, Pease, and Pearson in measuring the velocity of light in vacuum.

adjacent face of R and is observed at C by the eyepiece L_2 and the totally reflecting prism P .

It will be seen that the total light path with this arrangement is some ten miles. The exact distance between the faces of M_1 and M_2 was found by measuring the separation between marks on brass plates embedded in concrete piers beside the pipe. As in the previous determinations, R was rotated by an air blast at such a speed that one face just replaced the next in the time required for the light to return.

A plot of the results of 2885 determinations is shown in Fig. 19G. The broken curve is the curve expected for purely random errors having a probable error for one observation of 9 km/sec. The mean value 299,774 has a computed probable error of only 0.2 km/sec, but this apparently overestimates the true accuracy of the result.

19.8. Kerr-cell Method. Determinations by this method have equaled if not surpassed the accuracy of those by the rotating mirror. In 1925 Karolus and Mittelstaedt developed an improvement on Fizeau's method (Sec. 19.3) based on the use of the *electrooptic shutter*. This

device is capable of chopping a beam of light several hundred times more rapidly than can be done with a cogwheel. Hence a much shorter baseline can be used, and the entire apparatus can be contained in one building so that the atmospheric conditions are accurately known. Figure 19H(a) illustrates the electrooptic shutter, which consists of a Kerr cell K between two crossed nicol prisms N_1 and N_2 . K is a small

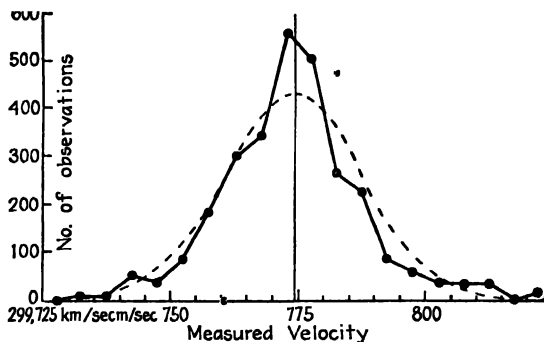


FIG. 19G. Error-distribution curve of nearly three thousand different determinations of the velocity of light in vacuum.

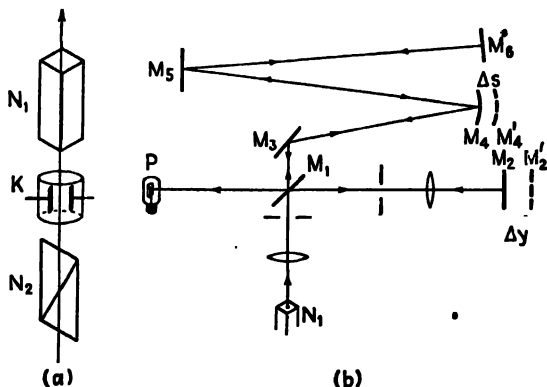


FIG. 19H. Anderson's method of measuring the velocity of light. (a) Electrooptic shutter. (b) The light paths.

glass container fitted with sealed-in metal electrodes and filled with pure nitrobenzene. Although the operation of this shutter depends on certain properties of polarized light to be discussed later (Chap. 29), all that need be known here in order to understand the method is that no light is transmitted by the system until a high voltage is applied to the electrodes of K . Thus by using an electrical oscillator which delivers a radio-frequency voltage, a light beam can be interrupted at the rate of many millions of times per second.

The first measurements based on this principle used two shutters, one for the outgoing and one for the returning light. Except for the shorter distances, the method closely resembled Fizeau's. Subsequent improvements have led to the apparatus shown in Fig. 19H(b), which was used by W. C. Anderson in 1941. To avoid the difficulty of matching the characteristics of two Kerr cells, he used only one, and divided the transmitted light pulses into two beams by means of the half-silvered mirror M_1 . One beam traversed the short path to M_2 and back through M_1 to the detector P . The other traveled a longer path to M_3 by reflections at M_3 , M_4 , and M_5 , then retraced its course to M_1 which reflected it to P as well. This detector P was a photomultiplier tube, which gave a strong response if the two sets of light pulses arrived in phase with each other but none if they arrived out of phase. The length l of these pulses is equal to the distance that light travels in a single period T of the oscillator which drives the Kerr cell.* A determination of l and T will thus give the velocity of light.

To measure l , Anderson arranged the mirrors so that the segment $M_4M_5M_3M_2M_1$ of the longer path was very nearly $11l$. This segment could be cut out of the path by substituting for the mirror M_4 another, M'_4 , which returned the light directly to M_3 . The length s of the segment was measured with an accurately calibrated invar tape to be 171.864 ± 0.002 m. The exact difference between this distance and $11l$ was then found as follows: With M'_4 in place, the path difference was adjusted by moving M_2 to a position where zero response was observed on the detector. Under this condition

$$N_1M_1M_3M'_4M_3M_1P - N_1M_1M_2M_1P = \frac{l}{2}$$

Next M_4 was substituted for M'_4 , and a new position was found for the movable mirror M_2 , say M'_2 , which again gave zero response. Then one has

$$N_1M_1M_3M_4M_5M_3M_4M_2M_1P - N_1M_1M'_2M_1P = 11\frac{1}{2}l$$

Subtracting the previous equation and employing the abbreviations $M_4M_5M_3M_2M_1 = s$, $M_4M'_4 = \Delta s$, and $M_2M'_2 = \Delta y$, one finds that

$$11l = s - 2\Delta s - 2\Delta y$$

The small shift Δs involved in interchanging the mirrors was measured by micrometer calipers as $2.477 \pm .012$ cm. Thus each determination

* Since the shutter transmits at each voltage peak, whether positive or negative, one would expect to use $\frac{1}{2}T$ here. Actually Anderson applied a d-c bias to the cell so that each cycle gave a single voltage maximum.

of Δy gave a value for l . The period T of the oscillator could be easily found to an accuracy of better than one part in a million. The frequency was held at 19.2 Mc/sec by a crystal control, and the latter was checked against the standard frequencies broadcast from Arlington.

The reader will see the resemblance of Anderson's apparatus to a Michelson interferometer for radio waves, since the light pulses have a length essentially equal to the wavelength of the radio waves given by the Kerr-cell oscillator. It is not exactly equal, however, because the velocity involved in the experiment is the group velocity of light in air and not the velocity of radio waves. In his final investigation, Anderson made a total of 2895 observations, and the observed velocities l/T , after correction to vacuum, yielded an average of $299,776 \pm 6$ km/sec. The agreement of this figure with the result of the most accurate work with rotating mirrors, described in the preceding section, is excellent.

TABLE 19I*

Date	Investigator	Method	Observed velocity c , km/sec
1875	Cornu	Rotating mirror	$299,990 \pm 200$
1880	Michelson	Rotating mirror	$299,910 \pm 50$
1883	Newcomb	Rotating mirror	$299,860 \pm 30$
1883	Michelson	Rotating mirror	$299,853 \pm 60$
1926	Michelson	Rotating mirror	$299,796 \pm 4$
1928	Mittelstaedt	Kerr cells	$299,778 \pm 10$
1932	Pease and Pearson	Rotating mirror	$299,774 \pm 2$
1941	Anderson	Kerr cell	$299,776 \pm 6$
1923	Mercier	Waves along wires	$299,782 \pm 30$
1906	Rosa and Dorsey	Ratio of electrostatic to electromagnetic units	$299,781 \pm 10$

* See R. T. Birge, *Nature*, **134**, 771, 1934.

19.9. Indirect Methods. One of the important indirect methods used in determining the velocity of light is that of measuring the velocity of electric waves along wires. This method involves sending high-frequency electrical waves along one of two adjacent and parallel wires and back along the other to produce standing waves. By measuring the distance between nodes and the frequency of the oscillations the velocity c can be calculated. Accurate determinations by Mercier* in 1923 gave a value which, corrected by Dorsey, was $c = 299,782 \pm 30$ km/sec in vacuum. This is in excellent agreement with the most accurate values obtained by direct methods (Table 19I).

A second indirect method is that of finding the ratio between the electrostatic and the electromagnetic units of electricity. The most accurate determinations of this ratio were made by Rosa and Dorsey in 1906. They give a result which, when corrected by Birge, is $c = 299,781 \pm 10$ km/sec. This also checks with the best determinations.

Recently there has been considerable discussion concerning the apparent possibility of a small periodic change in the velocity of light. A careful review by Birge* would tend to show no real variation of this kind. He suggests for the best value of the velocity of light $c = 299,776 \pm 4$ km/sec.

19.10. Velocity of Light in Stationary Matter. The first experiment to measure the velocity of light in a transparent substance much denser than air was performed in 1850 by Foucault. This was regarded as a

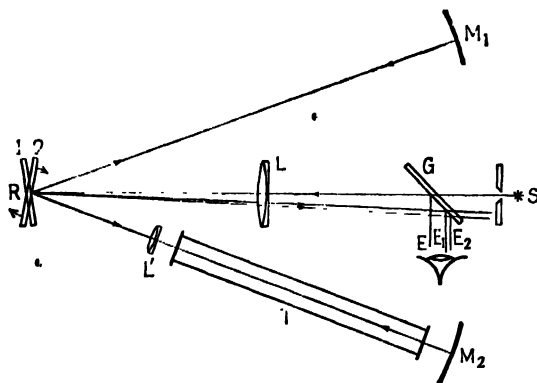


FIG. 19I. Foucault's apparatus for determining the velocity of light in stationary matter.

crucial experiment to decide between the corpuscular and wave theories of light. Newton's explanation of refraction by the corpuscular theory required that the corpuscles be attracted toward the surface of the denser medium, and therefore that they should travel faster in the medium. On the wave theory, however, it must be assumed that the light waves travel more slowly in the medium.

Foucault's apparatus for this experiment is shown in Fig. 19I. Light from a source S is reflected from the plane rotating mirror R to the two equidistant concave mirrors M_1 and M_2 . When R is in the position (1) the light travels to M_1 , back along the same path to R , through the lens L , and by reflection to the eye at E . When R is in the position (2) the light travels the lower path through an auxiliary lens L' and tube T to M_2 , back to R , through L to G and then to the eye at E . If now the tube T is filled with water and the mirror R is set into rotation, there will be displacement of the images from E to E_1 and E_2 . Foucault

observed that the light ray through the tube was displaced the most. This means that it took the light longer to travel the lower path through water than it did the upper path through air. The image observed was due to a fine wire parallel to and stretched across the slit. Since sharp images were desired at E_1 and E_2 , the auxiliary lens L' was necessary to avoid the effects of refraction at the ends of the tube T .

Much more accurate measurements were made by Michelson in 1885. Using white light, he found for the ratio of the velocity in air to that in water a value of 1.330. A denser medium, carbon disulfide, gave 1.758. In the latter case he noticed that the final image of the slit was spread out into a short spectrum, which could be explained by the fact that red light travels faster than blue light in the medium. The difference in velocity between "greenish blue" and "reddish orange" light was observed to be 1 or 2 per cent.

According to the wave theory of light, the index of refraction of a medium is equal to the ratio of the velocity of light in vacuum to that in the medium. If we compare the above figures with the corresponding indices of refraction for white light (water 1.334, carbon disulfide 1.635) we find that while the agreement is within the experimental error for water, the directly measured value is considerably higher than the index of refraction for carbon disulfide.

This discrepancy is readily explained by the fact that the index of refraction represents the ratio of the *wave velocities* in vacuum and in the medium ($n = c/v$), while the direct measurements give the *group velocities*. Now in a vacuum the two velocities become identical (Sec. 12.7) and equal to c , so that if we call the group velocity in the medium u , the ratios determined by Michelson were values of c/u , rather than c/v . The two velocities u and v are related by the general Eq. 12o

$$u = v - \lambda \frac{dv}{d\lambda}$$

The variation of v with λ may be found by studying the change of the index of refraction with color (Sec. 23.1), and it is found that v is greater for longer wavelengths, so that $dv/d\lambda$ is positive. Therefore u should be less than v , and this is precisely the result obtained above. Using reasonable values for λ and $dv/d\lambda$ for white light, the difference between the two values for carbon disulfide is in agreement with the theory to within the accuracy of the experiments. For water $dv/d\lambda$ is considerably smaller but nevertheless requires that the measured value of c/u should be 1.5 per cent higher than c/v . That this is not so indicates an appreciable error in Michelson's work. Recent work by R. A. Houstoun on the

velocity of light in water has given agreement not only as to the magnitude of the group velocity, but also as to its variation with wavelength.

At this point it should be emphasized that all the direct methods for measuring the velocity of light that we have described give the group velocity u and not the wave velocity v . Even though it is not evident in the aberration experiment that the wave is divided into groups, it should be obvious that since u is the velocity with which energy is transferred, and we always measure the energy, no direct measurement can give us v . In air the difference between u and v is small but nevertheless amounts to 2.2 km/sec. Michelson apparently did not apply this correction to his 1926 value, which should therefore have been quoted as $299,798 \pm 4$ km/sec.

19.11. Velocity of Light in Moving Matter. In 1859 Fizeau performed an important experiment to determine whether the velocity of light in a material medium is affected by motion of the medium relative to the

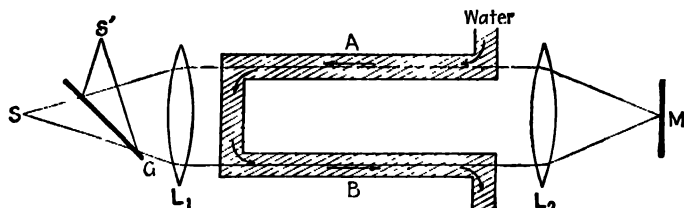


FIG. 19J. Fizeau's experiment for measuring the velocity of light in a moving medium.

source and observer. In Fig. 19J the light from S is split into two beams, and much the same way as in the Rayleigh refractometer (Sec. 13.16). The beams then pass through the tubes A and B containing water flowing rapidly in opposite directions. On reflection from M , the beams interchange so that when they reach L_1 one has traversed both B and A in the same direction as the flowing water while the other has traversed A and B in the opposite direction to the flow. The lens L_1 then brings the beams together to form interference fringes at S' .

If the light travels more slowly by one route than by the other, its optical path has effectively increased and a displacement of the fringes should occur. Using tubes 150 cm long and a water velocity of 700 cm/sec, Fizeau found a shift of 0.46 of a fringe when the direction of flow was reversed. This corresponds to an increase in the speed of light in one tube, and a decrease in the other, of about half of the velocity of the water.

This experiment was later repeated by Michelson with the improved apparatus shown in Fig. 19K. Here the light from S is divided into two beams by partial reflection at the surface of the half-silvered mirror

G. After traversing the tubes in which the water is circulated, as before, they are recombined to produce interference by the same mirror. Michelson observed a shift corresponding to an alteration of the velocity of light by 0.434 of the velocity of the water.

19.12. Fresnel Dragging Coefficient. The above results were compared with a formula derived by Fresnel in 1818, using the elastic-solid theory of the ether. On the assumption that the density of the ether in the medium is greater than that in vacuum in the ratio n^2 , he showed that the ether is effectively dragged along with a moving medium with a velocity

$$v' = v \left(1 - \frac{1}{n^2} \right) \quad (19b)$$

where v is the velocity of the medium, and n its index of refraction. For water, which has $n = 1.333$ for sodium light, this gives $v' = 0.437v$, in reasonable agreement with Michelson's value for white light quoted in

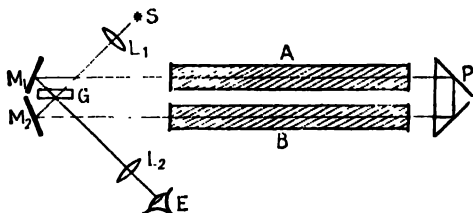


FIG. 19K. Michelson's experiment for measuring the velocity of light in a moving medium.

the previous paragraph. The fraction $1 - (1/n^2)$ will be referred to as *Fresnel's dragging coefficient*.

19.13. Airy's Experiment. An entirely different piece of experimental evidence shows that Fresnel's equation must be very nearly correct. In 1872 Airy remeasured the angle of aberration of light (Sec. 19.2), using a telescope filled with water. Upon referring to Fig. 19B(b), it will be seen that if the velocity of the light with respect to the solar system be made less by entering water, one would expect the angle of aberration to be increased. Actually the most careful measurements gave the same angle of aberration for a telescope filled with water as for one filled with air.

We may use this negative result to prove that the light must be dragged along by the water in the telescope with the velocity given by Eq. 19b. Here it is necessary to take account of the fact that, even if the telescope were at rest in the solar system, there would be in general a deflection of the light owing to refraction on entering the telescope.

LS. Applying the law of sines, that the three sides of any triangle are proportional to the sines of their opposite angles, we obtain from the triangle $L'QT$,

$$\frac{QT}{\sin \gamma} = \frac{L'Q}{\sin \beta}$$

or

$$\frac{v - v'}{\sin \gamma} = \frac{c/n}{\sin \left(\frac{\pi}{2} - \alpha \right)} = \frac{c}{n \cos \alpha}$$

Substituting the value $\sin \gamma = (\sin \alpha)/n$ from Eq. 19c, we have

$$\frac{n(v - v')}{\sin \alpha} = \frac{c}{n \cos \alpha}$$

Noting that $(\sin \alpha)/(\cos \alpha) = \tan \alpha = v/c$, we then find

$$\frac{v}{c} = \frac{n^2(v - v')}{c}$$

and

$$v' = v \left(1 - \frac{1}{n^2} \right)$$

which is identical with Eq. 19b. Therefore we see that, to explain Airy's result, we must assume a dragging coefficient agreeing with that found by direct measurements. The equation to be exact must still be corrected by a small factor depending on the dispersion of the medium.

19.14. Effect of Motion of the Observer. We have seen that in the phenomenon of aberration the apparent *direction* of the light reaching the observer is altered when he is in motion. One might therefore expect to be able to find an effect of such motion on the *magnitude* of the observed velocity of light. Referring back to Fig. 19B(b), we see that the apparent velocity $c' = v/\sin \alpha$ is slightly greater than the true velocity $c = v/\tan \alpha$. However, α is a very small angle, so that the difference between the sine and the tangent is much smaller than the error of measurement of α . A somewhat different experiment embodying the same principle has been devised, which should be sensitive enough to detect this slight change in the apparent velocity if it exists. Before describing this experiment, however, we consider in more detail the effect of motion of the observer on the apparent velocity of light.

In Fig. 19M, let the observer at O be moving toward B with a velocity v . Let an instantaneous flash of light be sent out at O . The wave will spread out in a circle with its center at O , and after 1 sec the radius of this circle will be numerically equal to the velocity of light c . But dur-

ing this time the observer will have moved a distance v from O to O' . Hence if the observer were in some way able to follow the progress of the wave, he would find an apparent velocity which would vary with the direction of observation. In the forward direction $O'B$ it would be $c - v$ and in the backward direction $O'A$, $c + v$. At right angles, in the direction $O'P$ he would observe a velocity $\sqrt{c^2 - v^2}$.

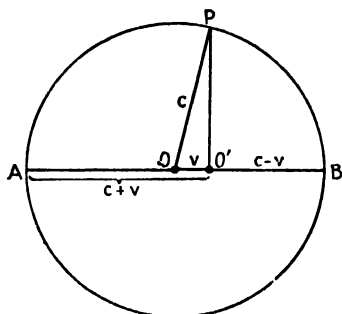


FIG. 19M. Velocity of light emitted by a moving source.

It is important to notice that in drawing Fig. 19M we have assumed that the velocity of the light was not affected by the fact that the source was in motion as it emitted the wave. This is to be expected for a wave which is set up in a stationary medium, as for instance a sound wave in the air. The hypothetical medium carrying light waves is the ether, and if v is the velocity with respect to the ether, the same result is expected. For an experiment performed in air, the Fresnel dragging coefficient $1 - (1/n^2)$ is so nearly zero that it may be neglected. Thus if the observer were moving with the velocity v of the earth in its orbit, these considerations lead us to expect the changes in the apparent velocity of light described above. Effectively the ether should be moving past the earth with a velocity v , and if any effects on the velocity of light were found, they could be said to be due to an ether wind or *ether drift*. It would not be surprising if this drift did not correspond to the velocity of the earth in its orbit, since we know that the solar system as a whole is moving toward the constellation Hercules with a velocity of 19 km/sec and it is more reasonable to expect the ether to be at rest with respect to the system of "fixed stars" than with respect to our solar system.

19.15. The Michelson-Morley Experiment. This experiment, perhaps the most famous of any experiment with light, was undertaken in 1881 to investigate the possible existence of an ether drift. In principle it consisted merely of observing whether there was any shift of the fringes in the Michelson interferometer when the instrument was turned through an angle of 90° . Thus in Fig. 19N let us assume that the

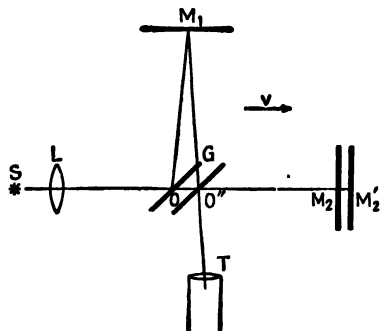


FIG. 19N. The Michelson interferometer as a test for ether drift.

interferometer is being carried along by the earth in the direction OM_1 , with a velocity v with respect to the ether. Let the mirrors M_1 and M_2 be adjusted for parallel light, and let $OM_1 = OM_2 = d$. The light leaving O in the forward direction will be reflected when the mirror is at M'_2 and will return when the half-silvered mirror G has moved to O'' . Using the expressions for the velocity derived in the previous section, the time required to travel the path OM'_2O'' will be

$$T_1 = \frac{d}{c+v} + \frac{d}{c-v} = \frac{2cd}{c^2 - v^2}$$

and the time to travel OM_1O'' will be

$$T_2 = \frac{2d}{\sqrt{c^2 - v^2}}$$

Each of these expressions may be expanded into series, giving

$$T_1 = \frac{2cd}{c^2 - v^2} = \frac{2d}{c} \left(1 + \frac{v^2}{c^2} + \frac{v^4}{c^4} + \dots \right) \cong \frac{2d}{c} \left(1 + \frac{v^2}{c^2} \right)$$

and

$$T_2 = \frac{2d}{\sqrt{c^2 - v^2}} = \frac{2d}{c} \left(1 + \frac{v^2}{2c^2} + \frac{3v^4}{4c^4} + \dots \right) \cong \frac{2d}{c} \left(1 + \frac{v^2}{2c^2} \right)$$

Thus the result of the motion of the interferometer is to increase both paths by a slight amount, the increase being twice as large in the direction of motion. The difference in time, which would be zero for a stationary interferometer, now becomes

$$T_1 - T_2 = \frac{2d}{c} \left(1 + \frac{v^2}{c^2} \right) - \frac{2d}{c} \left(1 + \frac{v^2}{2c^2} \right) = d \frac{v^2}{c^3}$$

To change this to path difference we multiply by c , obtaining

$$\Delta = d \frac{v^2}{c^2}$$

If now the interferometer is turned through 90° , the direction of v is unchanged, but the two paths in the interferometer will be interchanged. This would introduce a path difference Δ in the opposite sense to that obtained before. Hence we expect a shift corresponding to a change of path of $2dv^2/c^2$.

Michelson and Morley made the distance d large by reflecting the light back and forth between 16 mirrors as illustrated in Fig. 190. To avoid distortion of the instrument by strains, it was mounted on a large concrete block floating in mercury, and observations were made as it

was rotated slowly and continuously about a vertical axis. In one experiment d was 11 m, so that if we take $v = 18.6$ mi/sec and $c = 186,000$ mi/sec, we find a change in path of 2.2×10^{-5} cm. For light of wavelength 6×10^{-5} , this corresponds to a change of 0.4λ , so that the fringes should be displaced by two-fifths of a fringe. Careful observations showed that no shift occurred as great as 10 per cent of this predicted value.

This negative result, indicating the absence of an ether drift, was so surprising that the experiment has since been repeated with certain modifications by a number of different investigators. All have confirmed Michelson and Morley in showing that, if a real displacement of the

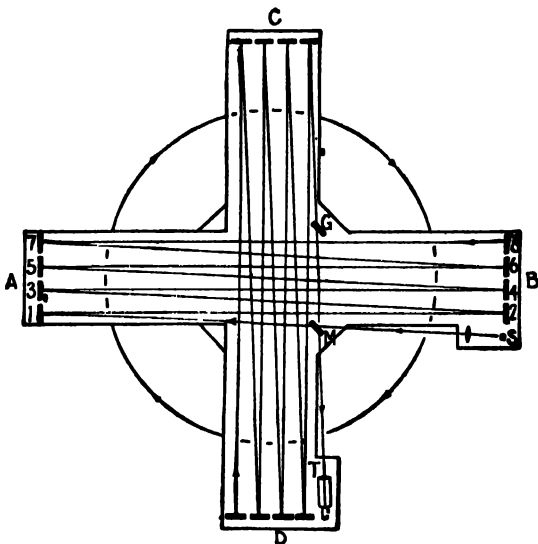


FIG. 190. Miller's arrangement of the Michelson-Morley experiment to detect ether drift.

fringes exists, it is at most but a small fraction of the expected value. The most extensive series of measurements has been made by D. C. Miller.* His apparatus is essentially that of Michelson and Morley (Fig. 190) but on a larger scale. With a light path of 64 m, Miller obtained evidence for a small shift of the fringes, which varies periodically with sidereal time. Although these shifts are only about one-twentieth of that expected for the earth's orbital motion through a stationary ether, Miller interprets them as indicating an absolute motion of the earth in space with a velocity of 208 km/sec superimposed on the orbital motion.

19.16. Principle of Relativity. The negative result obtained by Michelson and Morley, and by most of those who have repeated their experi-

ment, forms the basis of the restricted *theory of relativity*, put forward by Einstein* in 1905. The two fundamental postulates on which this theory is based are

1. *Principle of Relativity of Uniform Motion.* The laws of physics are the same for all systems having a uniform motion of translation with respect to one another. As a consequence of this, an observer in any one system cannot detect the motion of that system by any observations confined to the system,
2. *Principle of the Constancy of the Velocity of Light.* The velocity of light in any given frame of reference is independent of the velocity of the source. Combined with (1), this means that the velocity of light is independent of the relative velocity of the source and observer.

Returning to our illustration (Fig. 19M) of an observer who sends out a flash of light at O while moving with a velocity v , the above postulates would require that any measurements made by the observer at O' would show that he is the center of the spherical wave. But an observer at rest at O would find that he too is at the center of the wave. The reconciliation of these apparently contradictory statements lies in the fact that the space and time scales for the moving system are different from those for a fixed system. Events separated in space which are simultaneous to an observer at rest do not appear so to one moving with the system.

Another important consequence of Einstein's theory is that the length of an object in motion as measured by an observer at rest is shortened by an amount which increases with the velocity. If l_0 is the length of the object at rest, its length l when it is in motion with a velocity v becomes

$$l = l_0 \left(1 - \frac{v^2}{c^2} \right)^{\frac{1}{2}} \quad (19d)$$

c being the velocity of light. This is the celebrated Fitzgerald-Lorentz contraction, originally proposed to account for the negative result of the Michelson-Morley experiment. If the arm of the interferometer in the direction of motion becomes shorter by this amount, the difference in path due to ether drift is exactly canceled. Of course, such a contraction

could never be measured directly, since the length of the measuring stick would be shortened in the same proportion as the object to be measured.

Among the other predictions of the theory which have been verified by experiment is a variation of mass with velocity according to the equation

$$m = \frac{m_0}{\sqrt{1 - (v^2/c^2)}} \quad (19e)$$

This equation has been accurately verified by measuring the decrease in the ratio of charge to mass for an electron moving at high speed. The increase in mass can be shown to be related to the kinetic energy by the equation

$$\Delta m = \frac{\Delta E}{c^2} \quad (19f)$$

Here ΔE is the increase in energy due to kinetic energy. This indicates that mass is equivalent to energy, so that one may be changed into the other. Entirely independent proof of this relation has been obtained in recent experiments on atomic disintegration, where energy is created at the expense of mass, and vice versa.

In the years 1913-1917 Einstein extended his original theory, which applied only to systems moving with uniform velocity, to include the case of accelerated systems. This extension is called the *general theory of relativity*. Although the theory as a whole represents a fundamental change in our point of view with regard to the mechanics of moving systems, there are relatively few instances where it leads to results sufficiently at variance with the classical mechanics of Newton to be subject to experimental verification. Three predictions which have been satisfactorily verified are

1. *The Precession of the Perihelion of the Planet Mercury.* According to the restricted theory (Eq. 19e), the mass of the planet increases as it obtains a higher velocity at perihelion. Calculations on this basis gave a precession smaller than the observed value. The general theory, however, accounts quantitatively for the effect.
2. *The Deflection of Light Rays in Passing Close to the Sun.* The general theory predicts a bending of light rays by a gravitational field, resulting in an apparent outward displacement of the stars in the neighborhood of the sun. Earlier theories also required such a bending of light, but by just half the amount calculated by Einstein. The most recent measurements of plates taken during a total solar eclipse are in close agreement with the relativity theory.

3. *The Decrease in the Frequency of Light Emitted by a Source in a Strong Gravitational Field.* That such a decrease actually exists for the light from the sun was proved by St. John, who compared the wavelengths of the solar lines with those from laboratory sources. He found a barely detectable displacement toward longer wavelengths, as predicted by the theory. Further confirmation has been obtained from the spectra of the white dwarf stars, where the gravitational fields are many times stronger than on the sun.

These experimental proofs of the theory have been sufficiently convincing to lead to the general acceptance of the correctness of the theory of relativity. While the theory does not directly deny the existence of the ether postulated by Fresnel, it says very definitely that no experiment we can ever perform will prove its existence. For if it were possible to find the motion of a body with respect to the ether, we could regard the ether as a fixed coordinate system with respect to which all motions are to be referred. But it is one of the fundamental consequences of relativity that any coordinate system is equivalent to any other, and no one has any particular claim to finality. Thus, since a fixed ether is apparently not observable, there is no reason for retaining the concept. It cannot be denied, however, that it is historically important and that some of the most important advances in the study of light have come through the assumption of a material ether.

19.17. The Three First-order Relativity Effects. There are three optical effects the magnitude of which depends on the first power of v/c . They are

1. The Doppler effect
2. The aberration of light
3. The Fresnel dragging coefficient

Equations for these effects have been derived on the basis of classical theory in Secs. 11.6, 19.2, and 19.12. Now it is characteristic of the theory of relativity that it yields the same results for first-order effects as does the classical theory. Only in second-order effects, which depend on v^2/c^2 , do the predictions of the two theories differ. The Michelson-Morley experiment belongs to this class. Even for the first-order effects listed above, the results from the two theories differ in the small terms of the second and higher power of v/c . In the relativity theory, these equations are derived by applying the *Lorentz transformation*. This is a process of translating the description of a motion in terms of one system of coordinates into a description of the same motion in terms of another system which is in uniform motion with respect to the first. Although

it is not practicable to give the mathematics of this process here, we shall state the chief results and discuss them briefly.

The relativity equation for the Doppler effect was given in Sec. 11.6 as a power series for the purpose of comparison with the classical formulas. This series was an expansion of the closed formula

$$\nu' = \nu \frac{\sqrt{1 - (v^2/c^2)}}{1 - (v/c)} \quad (19g)$$

Here v is the relative velocity of approach of the source and observer along the line joining them. As was pointed out before, Eq. 19g yields a Doppler shift just halfway between that given by the two classical formulas, the one for the observer in motion, $\nu' = \nu \left(1 + \frac{v}{c}\right)$, and the

other for the source in motion, $\nu' = \nu \sqrt{1 - \frac{v^2}{c^2}}$. The first-order terms in the series expansion of all three of these expressions are the same, but they differ in the second- and higher-order terms (see Sec. 11.6). In Eq. 19g, the term $\sqrt{1 - (v^2/c^2)}$ is of second order, and theoretically originates from the fact that the rate of a moving clock is slower than that of a stationary one. Ives has given an elegant demonstration of this fact by comparing the frequency of the radiation emitted by hydrogen atoms in a high-speed beam moving first toward the spectroscope, then away from it. In addition to the large first-order shifts of the line toward higher and lower frequencies respectively in these two cases, he observed and measured a small additional shift which was toward lower frequencies in both cases. Since the term in question contains the square of the velocity, it will be the same for either sign of v . This experiment constitutes another verification of the theory of relativity by observation of a second-order effect which does not exist according to the classical theory. It might also be mentioned that relativity predicts a second-order Doppler shift even when the source is moving at right angles to the line of sight.

The interpretation of the aberration of light and of Airy's experiment is simpler from the relativistic point of view. According to the second fundamental postulate, the velocity of light must always be c to any observer, regardless of his motion. Hence, referring to Fig. 19B(b), the observed velocity labeled c' must now be labeled c . The formula for the angle of aberration, instead of being $\tan \alpha = v/c$, then becomes

$$\sin \alpha = \frac{v}{c} \quad (19h)$$

It is well known that the sine and the tangent differ only in respect to terms of the second order. Here the angle is so small that in all likelihood the difference will never be detected. In Airy's experiment, the expectation of observing an increase of the angle when the telescope was filled with water arose from the assumption that the water would decrease the velocity of the light with respect to the solar system, in which the ether was regarded as fixed. But according to the point of view of relativity the only "true" velocity of light is its velocity in the coordinate system of the observer, and this is inclined at the angle α given by Eq. 19*h*. Hence reducing the magnitude of this velocity by allowing the light to enter water will obviously make no change in its direction.

A positive effect corresponding to Fresnel's ether drag can be observed when the medium is in motion with respect to the observer (Sec. 19.12), but its interpretation by the theory of relativity is entirely different. One result of the Lorentz transformation is that two velocities in coordinate systems that are in relative motion do not add according to the methods used in classical mechanics. For example the resultant of two velocities in the same line is not their arithmetic sum. Let us call V_0 the velocity of light in the coordinate system of a moving medium, and v the velocity of this medium in the observer's coordinate system. Then the resultant velocity V of the light with respect to the observer, instead of being merely $V_0 + v$, must be taken as

$$V = \frac{V_0 + v}{1 + (V_0/c)(v/c)} \quad (19i)$$

The student can easily verify the fact that this equation gives the same velocity V for any observer in motion with the velocity v , in the case that $V_0 = c$; i.e., in a vacuum. The expression for the Fresnel dragging coefficient follows at once from Eq. 19*i*, if one neglects second-order terms. Thus the binomial expansion gives

$$\begin{aligned} V &= (V_0 + v) \left(1 - \frac{V_0}{c} \cdot \frac{v}{c} - \dots \right) \\ &= V_0 + v - \frac{V_0^2 v}{c^2} - \frac{v^2 V_0}{c^2} - \dots \end{aligned}$$

The last term is again a quantity of the second order and is to be neglected. Then we obtain, by substituting n for c/V_0 ,

$$V = \frac{c}{n} + v \left(1 - \frac{1}{n^2} \right) \quad (19j)$$

The velocity as seen by the observer is changed by the fraction $1 - (1/n^2)$, which is just the value required by Eq. 19b. No assumption of any "dragging" is involved in the relativity arguments, nor is the existence of an ether even postulated.

Problems

1. Derive the appropriate formula for the velocity of light in a Foucault rotating mirror apparatus arranged as shown in Fig. 19D. Let d = distance from rotating mirror R to fixed mirror M , r = distance from source or image to R , s = displacement of image due to rotating mirror, n = revolutions per second of mirror R , and V = velocity of light.

2. One of the largest of Jupiter's moons has an average period of 42 hr 28 min 16 sec. Assuming the radius of the earth's orbit to be 1.4967×10^8 km, and the velocity of light in vacuum to be 2.99778×10^8 km/sec, calculate (a) the number of revolutions of the moon which occur between two consecutive conjunctions of the earth and Jupiter, (b) the difference between the maximum and minimum observed periods of this moon. Assume Jupiter to have a period of 11.86 years.

3. (a) Determine the speed of the toothed wheel used in the actual determination of the velocity of light made by Fizeau, giving the speed in revolutions per second for the first three zeros of intensity. (b) In Cornu's repetition of the experiment, the base line was increased to 22.9 km. Repeat part (a) for this case, taking the diameter of the toothed wheel as 40 mm, and the number of teeth as 180.

4. For Fizeau's experiment, plot an intensity curve of the light reaching the observer's eye as a function of the speed of the wheel. Assume the opening between the teeth to be (a) equal to the width of the teeth, (b) two-thirds the width of the teeth, and (c) three-halves the width of the teeth. Assume a point source of light.

5. Calculate the speed of rotation of the mirror in Foucault's experiment, Fig. 19D. Assume the distance RM to equal 20 m, RS to equal 1 m, and the displacement EE' to be 0.7 mm, as it was in the actual experiment.

6. Assuming a Fresnel dragging coefficient of $\gamma = 0.445$ for water, find the velocity of the water stream necessary to produce a fringe shift of three fringes, upon reversal of the water stream, for light of wavelength 6870 Å. Assume Fizeau's experimental arrangement illustrated in Fig. 19J, with the tubes A and B each 5 m long.

7. Solve Prob. 6 in the case $\gamma = 0.465$ and $\lambda = 4500$ Å.

8. Calculate the Fresnel dragging coefficient for air at normal atmospheric pressure and temperature.

9. In an experimental determination of the Fresnel dragging coefficient by Michelson's method, assume that the length of each tube was 150 cm and the velocity of the water 7 m/sec. If light of $\lambda = 5461$ Å in vacuum were used, for which the refractive index of water $n = 1.33447$ and $dn/d\lambda = -383 \text{ cm}^{-1}$, find (a) the value of the dragging coefficient γ neglecting the effect of dispersion, (b) its value including dispersion, which increases γ by adding a term $-(\lambda/n)(dn/d\lambda)$, and (c) an explicit formula for the fringe shift.

10. Derive a formula for the velocity of light in the rotating mirror apparatus as originally used by Fizeau, where the lens was placed between the source and the rotating mirror and the distant fixed mirror was at the conjugate focus of the lens. Let n = frequency of rotating mirror, d = distance between mirrors R and M ,

u = distance from source to lens (object distance), l = distance from lens to rotating mirror, s = displacement of image, and V = velocity of light.

11. In Anderson's Kerr-cell determination of the velocity of light, a correction of 84 km/sec was added to the value measured in air to obtain the value $c = 299,776$ km/sec for the velocity in vacuum. Assuming the distances s , Δs , and the period T given in Sec. 19.8, compute the corresponding value of Δy .

12. Calculate the energy in ergs equivalent to the mass of an electron (see Eq. 19f).

13. Find the mass of an electron moving with $\frac{1}{2}$, $\frac{1}{3}$, $\frac{2}{3}$, $\frac{4}{5}$, and $\frac{9}{10}$ of the velocity of light.

CHAPTER 20

THE ELECTROMAGNETIC CHARACTER OF LIGHT

Our study of the properties of light has thus far led us to the conclusion that light is a wave motion, propagated with an extremely high velocity. In the explanation of interference and diffraction it was not necessary to make any assumption as to the nature of the displacement y that appears in our wave equations. This is because in these subjects we were concerned only with the interaction of light waves with each other. In the succeeding chapters we are to consider subjects in which the interaction of light with matter plays a part, and here it becomes necessary to specify the physical nature of the quantity y , which is usually termed the *light vector*. Fresnel, who in 1814 first gave the satisfactory explanation of interference and diffraction by the wave theory, imagined the light vector to represent an actual displacement of a material ether, which was conceived as an all-pervading substance of very small density and of high rigidity. This "elastic-solid" theory had considerable success in interpreting optical phenomena and was strongly supported by many leading investigators in the field, such as Lord Kelvin, as late as 1880.

20.1. Transverse Nature of Light Vibrations. The principal objection to the elastic-solid theory lay in the fact that light had been proved to be exclusively a transverse wave motion, *i.e.*, the vibrations are always perpendicular to the direction of motion of the waves. No longitudinal waves of light have ever been detected. The experimental evidence for this comes from the study of the polarization of light (Chap. 24) and is perfectly definite, so that we may here take the fact as established. Now all elastic solids with which we are familiar are capable of transmitting longitudinal as well as transverse waves; in fact, under some circumstances it is impossible to set up a transverse wave without at the same time starting a longitudinal one. Many suggestions were made to overcome this difficulty, but all were highly artificial. Furthermore, the idea of a material ether itself seemed rather forced, inasmuch as its remarkable properties could not be detected by ordinary mechanical experiments.

Thus the time was ripe when Maxwell* proposed a theory which not

* J. Clerk Maxwell (1831-1879). Professor of experimental physics at Cambridge University, England. Contributed a paper to the Royal Society at the age of fifteen.

only *required* the vibrations of light to be strictly transverse but also gave a definite connection between light and electricity. In a paper read before the Royal Society in 1864, entitled "A Dynamical Theory of the Electromagnetic Field," Maxwell expressed the results of his theoretical investigations in the form of four fundamental equations which have since become famous as *Maxwell's equations*. They were based on the earlier experimental researches of Oersted, Faraday, and Joseph Henry concerning the relations between electricity and magnetism. They summarize these relations in concise mathematical form, and constitute a starting point for the investigation of all electromagnetic phenomena. We shall show in the following sections how they account for the transverse waves of light.

20.2. Maxwell's Equations for a Vacuum. The derivation of these equations will not be given here, since it would involve a rather extensive review of the principles of electricity and magnetism.* Instead we shall in this chapter merely state the equations in their simplest form, applicable to empty space, and then prove that they predict the existence of waves having the properties of light waves. The modifications that must be introduced in dealing with different kinds of material media will be considered at the appropriate places in the following chapters.

Maxwell's equations may be written as four vector equations, but for those unfamiliar with vector notation we shall express them by differential equations. In this form the first two equations must be expressed by two sets of three equations each. For a vacuum these become, using a right-handed set of coordinates,

$$\left. \begin{aligned} \frac{1}{c} \frac{\partial E_x}{\partial t} &= \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} \\ \frac{1}{c} \frac{\partial E_y}{\partial t} &= \frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} \\ \frac{1}{c} \frac{\partial E_z}{\partial t} &= \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} \end{aligned} \right\} \quad (20a)$$

$$\left. \begin{aligned} -\frac{1}{c} \frac{\partial H_x}{\partial t} &= \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} \\ -\frac{1}{c} \frac{\partial H_y}{\partial t} &= \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} \\ -\frac{1}{c} \frac{\partial H_z}{\partial t} &= \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} \end{aligned} \right\} \quad (20b)$$

The other two equations may be written

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0 \quad (20c)$$

$$\frac{\partial H_x}{\partial x} + \frac{\partial H_y}{\partial y} + \frac{\partial H_z}{\partial z} = 0 \quad (20d)$$

Much of his work on the electromagnetic theory was accomplished while an undergraduate at Cambridge. His investigations in many fields of physics bear the stamp of genius. The kinetic theory of gases was given a solid mathematical foundation by Maxwell, whose name is associated with the well-known law of distribution of molecular velocities.

These partial differential equations give the relations in space and time between the vector quantities \mathbf{E} , the electric field strength, and \mathbf{H} , the magnetic field strength. Thus E_x , E_y , and E_z are the components of \mathbf{E} along the three rectangular axes x , y , and z , and H_x , H_y , H_z are the corresponding components of \mathbf{H} . The constant c is the ratio of the magnitudes of the electromagnetic and electrostatic units of current.

Equation 20c merely expresses the fact that no free electric charges exist in a vacuum. The impossibility of a free magnetic pole gives rise to Eq. 20d. Equations 20b express Faraday's law of induced electromotive force. Thus the quantities occurring on the left side of these equations represent the time rate of change of the magnetic field, and the spacial distribution of the resulting electric fields occurs on the right side. These equations do not give directly the magnitude of the emf, but only the rates of change of the electric field along the three axes. In particular problems the equations must be integrated to obtain the emf itself.

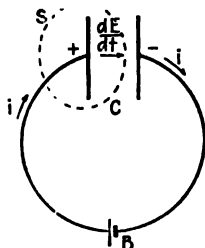


FIG. 20A. Illustrating the concept of displacement current.

20.3. Displacement Current. Maxwell's principal new contribution in giving these equations was the statement of Eqs. 20a. These come from an extension of Ampère's law for the magnetic field due to an electric current. The right-hand members give the distribution of the magnetic field H in space, but the quantities on the left side do not at first sight seem to have anything to do with electric current. They represent the time rate of change of the electric field. But Maxwell regarded this as the equivalent of a current, the *displacement current*, which flows as long as the electric field is changing and which produces the same magnetic effects as an ordinary conduction current.

One way of illustrating the equivalence of $\partial \mathbf{E} / \partial t$ to an electric current is shown in Fig. 20A. Imagine an electric condenser C to be connected to a battery B by conducting wires, the whole apparatus being in a vacuum with a vacuum between the condenser plates. As the current i flows for an instant, electric charge accumulates on the plates until the condenser is fully charged to the voltage of the battery. Through the closed surface S , a certain current has been flowing in during this instant, but none has apparently been flowing out. By considerations of continuity, Maxwell was led to assume that as much current should flow out of such a surface as flows in. But no current of the ordinary sort is flowing between the plates of the condenser. The condition of continuity can be satisfied only by regarding the change of the electric field

in this space as the equivalent of a displacement current, the current density j of which is proportional to $\partial E/\partial t$. In our system of units this current is given by $j = 1/4\pi$ times $\partial E/\partial t$. It will be noticed that the displacement current "flows" in a vacuum, but stops as soon as E becomes constant.

One sees at once the analogy between Eqs. 20*b* and 20*a*. By Eqs. 20*b* a changing magnetic field produces an emf. This was observed by Faraday and is very simple to verify experimentally. By Eqs. 20*a* a changing electric field should produce a magnetic field ("magnetomotive force"). This is a much less familiar idea and cannot be proved by any simple experiment. The reason for the difference is that no substance conducts magnetism as a wire conducts electricity. The peculiarity that some substances possess of being conductors for electricity is the only reason why Eqs. 20*b* were discovered before Eqs. 20*a*. The proof of the correctness of Eqs. 20*a* lies in the remarkable success of Maxwell's equations in accounting for phenomena of nature. It should be noted that Maxwell's equations 20*a* and 20*b* may be written in terms of the displacement current j by replacing the x component $(1/c)(\partial E_x/\partial t)$ by $4\pi j_x$ and the other components by similar expressions.

20.4. The Equations for Plane Electromagnetic Waves. Consider the case of plane waves traveling in the x direction, so that the wave fronts are planes parallel to the y, z plane. If the vibrations are to be represented by variations of E and H , we see that in any one wave front they must be constant over the whole plane at any instant, and their partial derivatives with respect to y and z must be zero. Therefore Eqs. 20*a* to 20*d* take the form

$$\left. \begin{aligned} \frac{1}{c} \frac{\partial E_x}{\partial t} &= 0 \\ \frac{1}{c} \frac{\partial E_y}{\partial t} &= -\frac{\partial H_z}{\partial x} \\ \frac{1}{c} \frac{\partial E_z}{\partial t} &= \frac{\partial H_y}{\partial x} \end{aligned} \right\} \quad (20e)$$

$$\frac{\partial H_x}{\partial x} = 0 \quad (20g)$$

$$\left. \begin{aligned} -\frac{1}{c} \frac{\partial H_x}{\partial t} &= 0 \\ -\frac{1}{c} \frac{\partial H_y}{\partial t} &= -\frac{\partial E_z}{\partial x} \\ -\frac{1}{c} \frac{\partial H_z}{\partial t} &= \frac{\partial E_y}{\partial x} \end{aligned} \right\} \quad (20f)$$

$$\frac{\partial E_x}{\partial x} = 0 \quad (20h)$$

Considering the first equation of Eqs. 20*e* and Eq. 20*h* together, it appears that the longitudinal component E_x is constant in both space and time. Similarly from the top line of Eqs. 20*f* and from Eq. 20*g*, H_x is also constant. These components can therefore have nothing to do with the wave motion, but must represent constant fields superimposed on the system of waves. For the waves themselves, we may therefore write

$$E_x = 0 \quad H_x = 0$$

This means, of course, that the waves are transverse, as stated above.

Of the four remaining equations, we see that the second equation 20e and the third equation 20f involve E_y and H_x , while the third equation 20e and the second equation 20f involve E_x and H_y . Let us assume, for example, that E_y represents the light vector, so that we are dealing with a plane-polarized wave with vibrations in the y direction. We should then have to put $E_x = H_y = 0$, and consider the two remaining equations

$$\begin{aligned}\frac{1}{c} \frac{\partial E_y}{\partial t} &= -\frac{\partial H_x}{\partial x} \\ \frac{1}{c} \frac{\partial H_x}{\partial t} &= \frac{\partial E_y}{\partial x}\end{aligned}\quad (20i)$$

We now differentiate the first equation with respect to t and the second with respect to x . This gives

$$\begin{aligned}\frac{1}{c} \frac{\partial^2 E_y}{\partial t^2} &= -\frac{\partial^2 H_x}{\partial x \partial t} \\ -\frac{1}{c} \frac{\partial^2 H_x}{\partial t \partial x} &= \frac{\partial^2 E_y}{\partial x^2}\end{aligned}$$

Eliminating the derivatives of H_x , we find

$$\frac{\partial^2 E_y}{\partial t^2} = c^2 \frac{\partial^2 E_y}{\partial x^2} \quad (20j)$$

In a similar way, by differentiation of the first equation 20i with respect to x and the second with respect to t , we find

$$\frac{\partial^2 H_x}{\partial t^2} = c^2 \frac{\partial^2 H_x}{\partial x^2} \quad (20k)$$

It remains to show that Eqs. 20j and 20k are the differential equations of wave motions traveling along the x axis. In Chap. 11, the most general equation for plane waves was Eq. 11d, which for the electric vector would be written

$$E_y = f(x - vt) \quad (20l)$$

That this is a solution of Eq. 20j may be shown by differentiating it twice with respect to the time. The first differentiation gives

$$\frac{\partial E_y}{\partial t} = -vf'(x - vt)$$

where f' is the derivative of the function f with respect to x . The second gives

$$\frac{\partial^2 E_y}{\partial t^2} = v^2 f''(x - vt)$$

f'' denoting the second derivative. But differentiation of Eq. 20l twice with respect to x gives

$$\frac{\partial^2 E_y}{\partial x^2} = f''(x - vt)$$

and comparing the last two equations,

$$\frac{\partial^2 E_y}{\partial t^2} = v^2 \frac{\partial^2 E_y}{\partial x^2} \quad (20m)$$

This becomes identical with Eq. 20j if we put

$$v = c \quad (20n)$$

The same velocity may be obtained for the waves of H_z , by comparison with Eq. 20k. Thus we see that two of the four equations in Eqs. 20e and 20f predict the existence of a wave of the electric vector, plane-polarized in the x, y plane, and an accompanying wave of the magnetic vector, plane-polarized in the x, z plane. The two waves are interdependent; neither can exist without the other. Both are transverse waves, and are propagated in a vacuum with the velocity c , the ratio of the electrical units (Sec. 20.2).

If we had started with the other two equations in Eqs. 20e and 20f, we would have obtained another pair of waves, plane-polarized with the electric vector in the x, z plane. This pair is quite independent of the other, and can exist separately from the other pair. A mixture of the two pairs, vibrating at right angles to each other, and with no constant phase relation between E_y and E_z , represents unpolarized light.

20.5. Pictorial Representation of an Electromagnetic Wave. The simplest type of electromagnetic wave is one in which the function f in Eq. 20l is a sine or cosine. This is a plane-polarized monochromatic plane wave. The three components of \mathbf{E} , and the three of \mathbf{H} , may for such a wave be written

$$\begin{aligned} E_x &= 0, & E_y &= r \cos \frac{2\pi}{\lambda} (x - ct), & E_z &= 0 \\ H_x &= 0, & H_y &= 0, & H_z &= r \cos \frac{2\pi}{\lambda} (x - ct) \end{aligned} \quad (20o)$$

By substituting the derivatives of these quantities in Eqs. 20a to 20d, it is easily verified that they represent a solution of Maxwell's equations.

Figure 20B shows a plot of the values of E_y and H_z along the x axis, according to Eq. 20o when $t = 0$. In a set of plane waves the values of E_y and H_z at any particular value of x are the same all over the plane

$x = \text{constant}$, so this figure merely represents the conditions for one particular value of y and z .

Two important points are to be noticed about Fig. 20B. In the first place, the electric and magnetic components of the wave are *in phase* with each other; *i.e.*, when E_y has its maximum value, H_z is also a maximum. The relative directions of these two vectors, as indicated in the figure, agree with Eqs. 20o. The second point to be noted is that the amplitudes of the electric and magnetic vectors are equal. That these two are numerically equal in the system of units used here is shown by the fact that, in Eqs. 20o, r is the amplitude of each wave.

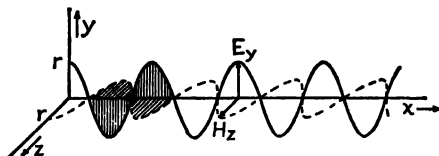


FIG. 20B. Distribution of the electric and magnetic vectors in a plane-polarized monochromatic wave.

20.6. Light Vector in an Electromagnetic Wave. The dual character of the electromagnetic wave raises the question as to whether it is the electric vector or the magnetic vector which is to be the light vector. This question has little meaning, since we could assume either one to represent the “displacements” we have been using in previous chapters. In every interference or diffraction phenomenon, the electric waves will mutually influence each other in exactly the same way as the magnetic waves. In one respect, however, the electric component plays a dominant part. It will be proved in Chap. 28 that it is the electric vector that affects the photographic plate and causes fluorescent effects. Presumably also the electric vector is the one that affects the retina of the eye. In this sense, therefore, the electric wave is the part that really constitutes “light,” and the magnetic wave, though no less real, is less important.

20.7. Energy and Intensity of an Electromagnetic Wave. The intensity of mechanical waves was shown in Chap. 11 to be proportional to the square of the amplitude. The same result follows from the electromagnetic equations. It can be shown that for an electromagnetic wave in a vacuum, the energy of the electric wave per unit volume is $E_0^2/8\pi$, where E_0 is the amplitude of the electric wave. Similarly $H_0^2/8\pi$ is the energy per unit volume in the magnetic wave. But since $E_0 = H_0 = r$, we have

$$\text{Energy per unit volume} = \frac{r^2}{4\pi} \quad (20p)$$

for a wave in vacuum. Half the energy of the wave is associated with the electric vector and half with the magnetic vector. In order to obtain the intensity, we have merely to multiply the above quantity by c , the velocity of the waves, since this will give the volume of the waves streaming through an area of unit cross section per second. The intensity is consequently also proportional to the square of the amplitude of either the magnetic or the electric wave. In a material medium, we shall see (Sec. 23.8) that the magnitudes of E and H are no longer equal. Nevertheless, it will appear that the above statement about the intensity still holds.

20.8. Radiation from an Accelerated Charge. A convenient method of representing an electric or magnetic field is by the use of lines of force. These are familiar to anyone who has studied elementary electricity and

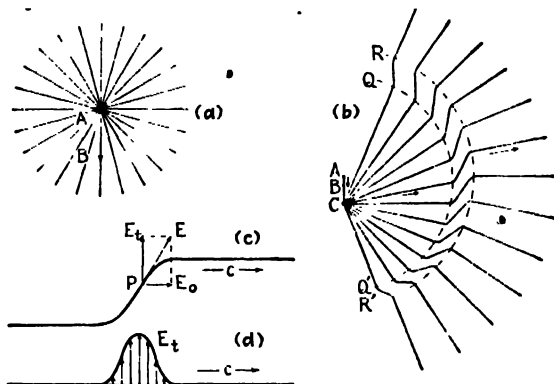


FIG. 20C. Emission of an electromagnetic pulse from an accelerated charge.

magnetism. Each line of force indicates the direction of the field at every point along the line, and this is such that a tangent to the line of force at any point gives the direction of the force on a small charge or pole placed at that point. That is, this tangent gives the direction of the electric or magnetic field at that point.

Consider a small positive electric charge at rest at the point A [Fig. 20C(a)]. The lines of force are straight lines diverging in every direction from the charge and are uniformly distributed in space. The same picture would hold if the charge were moving in the direction AB with the constant velocity v , assuming v to be not too large. In these two cases—charge at rest and charge in uniform motion—there is no radiation of electromagnetic waves.

In order to produce electromagnetic radiation, it is necessary to have *acceleration* of the charge. A particularly simple case is represented in Fig. 20C(b). Let the charge, originally at rest at A, be accelerated in

the direction AC . The acceleration a lasts only until the charge reaches the point B , and from that point on the charge moves with a constant velocity v . In this case we may obtain some information about the form of the lines of force radiating from this point. Let the time of the acceleration from A to B be Δt , and let the time of the uniform motion from B to C be t . When the charge has reached C , at a time $t + \Delta t$ after it starts, the parts of the original lines of force lying beyond the arc RR' , drawn about A with the radius $c(t + \Delta t)$, cannot have been disturbed in any way. This follows from the fact that any electromagnetic disturbance is propagated with the velocity c . At the point C the velocity is uniform and the lines of force as far as the arc QQ' , drawn about B with the radius ct , must be uniform and straight, since the charge has had a uniform velocity during the time t . Consequently we see that in order to have continuous lines of force they must be connected through the region between RR' and QQ' somewhat as shown in the figure. This gives a pronounced "kink" in each line. The exact form of the kink will depend upon the type of acceleration existing between A and B , *i.e.*, whether it is uniform or some type of nonuniform acceleration.

What is the significance of such a kink in a line of force? If we select some point P lying on the kink [Fig. 20C(c)], the vector E drawn tangent to the line at P gives the actual direction of the field at that point. This may be regarded as the resultant of the field E_0 which would be produced by the charge at rest, and a *transverse* field E_t . It is the vector E_t which represents the electric vector of the electromagnetic wave, referred to in the foregoing sections. If we carry out this construction for various points along the kink, we obtain the variations indicated in Fig. 20C(d). This is obviously not a periodic wave form, but merely a pulse. There will be a similar pulse of the H vector at right angles to E_t .

Several important features about the production of electromagnetic radiation are illustrated by this example. Most important is the fact that E_t exists only when the charge is *accelerated*. No radiation is produced if there is no acceleration of charge, and, conversely, an accelerated charge will always radiate to a greater or less extent. Also, the example shows how the electric field of the radiation can be transverse to the direction of propagation. The magnitude of the vector E_t obtained by the construction of Fig. 20C(d), *i.e.*, the amplitude of the wave, obviously depends on the steepness of the kink, and this is determined by how rapidly the charge was accelerated from A to B . It can be shown theoretically that the rate of radiation of energy from an accelerated charge is proportional to the square of the acceleration. Finally, we also find that the amplitude of the radiation varies with angle in such a way that

it is a maximum in directions perpendicular to the line AC and falls to zero in both directions along AC . The amplitude is easily shown to be proportional to the sine of the angle between AC and the direction considered.

20.9. Radiation from a Charge in Periodic Motion. If the charge in Fig. 20C, instead of undergoing a single acceleration, is caused to execute a periodic motion, the radiation will be in the form of continuous waves instead of a single isolated pulse. Any periodic motion involves accelerations, and hence will cause the charge to radiate. We shall here consider only two especially simple cases, that of linear simple periodic motion and that of uniform circular motion. If the positive charge of Fig. 20D(a) is moved with simple periodic motion between the limits A and B ,

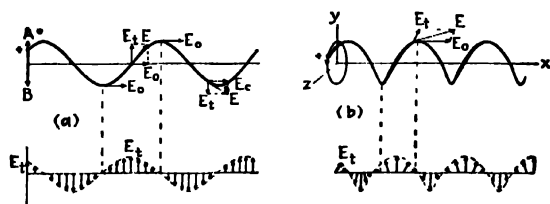


FIG. 20D. Emission of electromagnetic waves from a charge in periodic motion.

any line of force will be bent into the form of a sine curve. Let the upper curve Fig. 20D(a) represent one such line, say the one running out perpendicular to AB . At the particular instant shown, the electric force E at various points along the line has the direction of the tangent at those points. Resolving it into the undisturbed field E_o and the transverse component E_t , we find the various values E_t shown just below. These also take the form of a sine curve and represent the variation of the electric vector along the wave sent out. This is a plane-polarized wave.

In part (b) of the figure, the positive charge is revolving counterclockwise in a circle, in the y,z plane shown in perspective. The same construction now gives values of E_t which are constant in magnitude, but vary in direction along the wave. The heads of the arrows lie on a spiral similar to that of the line of force, but displaced one-quarter of a wavelength along the direction of propagation, which here is the x axis. This screwlike arrangement of the vectors is characteristic of a circularly polarized wave. It is worth pointing out here that, if the radiation along the y or z axes were examined, it would be found to be plane-polarized in the y,z plane. Actual observation of these two cases is possible in the Zeeman effect (Sec. 29.1).

20.10. Hertz's* Verification of the Existence of Electromagnetic Waves. We have seen that, starting with a set of equations governing the phenomena of electromagnetism, Maxwell was able to show the possibility of electromagnetic waves and to make definite statements about the production and properties of the waves. Thus he could say that they are generated by any accelerated charge, that they are transverse waves, and that they travel with the velocity c in free space. The experimental production and detection of the waves predicted by Maxwell were not long in forthcoming. In 1887, Heinrich Hertz began a remarkable series of experiments which constitute the first important experiments on radio waves, *i.e.*, electromagnetic waves of long wavelength. The essential features of Hertz's method are illustrated in Fig. 20E. Two plane brass plates are connected to a spark gap SG and

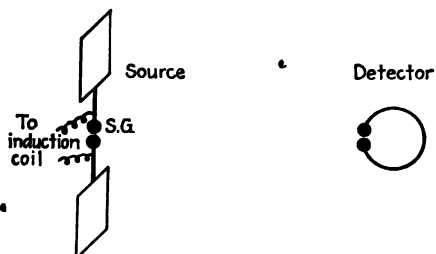


FIG. 20E. Showing source and detector of electromagnetic waves used by Hertz.

sparks are caused to jump across the gap by charging the plates to high voltage with an induction coil. It is known that the discharge of the plates by the spark is an oscillatory one. Each time the potential difference between the knobs of the gap reaches the point where the air in the gap becomes conducting, a spark passes. This represents a sudden surge of electrons across the gap, and the signs of the charges on the two plates become reversed. But since the air is still conducting, this will produce a return surge, another reversal of sign, and the process repeats until the energy is dissipated as heat by the resistance of the gap. The frequency of these oscillations depends on the inductance and capacity of the circuit. These were very small for Hertz's oscillator, and the frequency correspondingly high. In some of his experiments it reached 10^9 per sec. Thus we have an electric charge undergoing very rapid accelerations, and electromagnetic waves should be radiated.

* Heinrich Hertz (1857-1894). These experiments were carried on while he was professor of physics at Polytechnic Institute at Karlsruhe, Germany, in 1885-1889. He was then given a professorship at the University of Bonn, which he held until his untimely death.

In Hertz's experiment the presence of electromagnetic waves was detected at some distance from the oscillator by a resonating circuit consisting of a circular wire broken by a very narrow spark gap of adjustable length. The changing magnetic field in the wave induced an alternating emf in the circular wire, whose dimensions were such that the natural frequency of its oscillations was the same as that of the source. Thus the induced oscillations built up by resonance in the detector until they were sufficient to cause sparks to jump the gap.

It was a simple matter to show that the waves were plane-polarized with E in the y direction and H in the z direction. If the loop was turned through 90° so that it lay in the x,z plane, the sparks ceased to occur. Hertz performed many other experiments with these waves, showing among other things that the waves could be reflected and focused by curved metal reflectors, and that they could be refracted in passing through a large 30° prism of pitch. In these respects they therefore showed the same behavior as light waves.

20.11. Velocity of Electromagnetic Waves in Free Space. The most convincing proof of the reality of Hertz's electromagnetic waves lay in the demonstration that their velocity was that predicted by the theoretical equation (Eq. 20n). The velocity was measured not directly but indirectly by measuring the wavelength. Then from the known frequency of the oscillations the velocity could be found by the relation $v = \nu\lambda$. To measure the wavelength, standing waves were produced by interference of the direct waves with those reflected from a plane metal reflector. The positions of the nodes could be located by the fact that the detector ceased to spark at these points. With a frequency of 5.5×10^7 per sec, λ was found to be about 5.4 m, which gives v very close to 3×10^{10} cm/sec. The determination could not be made accurately, because the oscillations were highly damped, only three or four occurring after each spark, and the wavelength was therefore not accurately defined. More recent work by Mercier with undamped waves produced by a vacuum-tube oscillator, has given the result 2.9978×10^{10} cm/sec. It seems likely that with modern techniques of cavity resonators it may be possible before long to add another significant figure to the velocity of light.

According to Eq. 20n, this observed velocity should equal c , the ratio of the emu to the esu of current. This ratio has been accurately measured by different methods, the most recent value being 2.99781×10^{10} cm/sec. But this is just the measured velocity of electromagnetic waves and also agrees exactly with the latest measurements of the velocity of light by Michelson and others (Chap. 19). For air or other gases at atmospheric pressure, a slight modification in the equations is necessary

(Chap. 23), but the predicted velocity differs only slightly from that in vacuum.

Hence we are forced to conclude that light consists of electromagnetic waves of extremely short wavelength. Beside the evidence of polarization, which proves that light waves are transverse waves, there is much other evidence of this identity. Spectroscopy has shown that the atoms contain electrons and that by assuming the acceleration of these electrons as they move in orbits around the nucleus one can account for the polarization and intensity of the spectrum lines. Furthermore, as mentioned in Chap. 11, it has been shown that radio waves, which are obviously electromagnetic in character, join continuously on to the region of infrared light waves. Thus the explanation of light waves as an electromagnetic phenomenon, which in the hands of Maxwell was merely a very elegant theory, has since proved to be a reality, and we accept the electromagnetic character of light as an established fact. In treating the interactions of light with matter we shall therefore use the fact that light consists of oscillations of an electric field at right angles to the direction of propagation of the waves, accompanied by oscillations of the magnetic field, also at right angles to this direction and to the direction of the electric field.

Problems

1. The solutions of Maxwell's equations representing plane waves traveling in any arbitrary direction specified by the cosines l , m , and n of the angles that it makes with the x , y , and z axes, respectively, are $E = f(s - ct)$ and $H = f(s - ct)$, where $s = lx + my + nz$. By substituting these solutions in Eqs. 20a and 20b, prove that H is perpendicular to E . Show also that H and E are perpendicular to the direction of propagation, the latter being given by the right-hand screw rule, turning E towards H .

2. Assuming an electromagnetic wave traveling in the x direction and with the electric vector in the z direction, show that Maxwell's equations lead to Eq. 20j, with E_z replaced by E_x .

3. Prove that $E_y = f(x + ct)$, representing a wave traveling in the negative x direction, is also a solution of Eq. 20j.

4. Compare the force exerted on an electron by the electric and magnetic fields of the light wave. Assume the electron to be moving with a velocity of 10^7 cm/sec (roughly the orbital velocity in an atom) and to be acted on by sodium light of intensity 0.001 watt/cm² (about that from a sodium arc 2 m away). (NOTE: In Gaussian units, for which the equations of this chapter are written, E and H are measured in esu and emu respectively.)

5. An oscillator of frequency 3.5 Mc/sec is set up near a plane metal reflector, and the distance between the adjacent nodes in the standing waves is found to be 4.28 cm. What does this give for the velocity of light?

6. Make a sketch of the electromagnetic wave of Fig. 20B at a time one-quarter of a period later.

7. Prove by direct differentiation that Eqs. 20c represent a solution of Eqs. 20a to 20d.

8. Since the energy in an electromagnetic wave is equivalent to a certain mass by Einstein's relation, Eq. 19f, light waves possess momentum and exert pressure on a surface which absorbs or reflects them. Calculate the radiation pressure on a perfectly reflecting surface placed normal to full sunlight (0.13 watt/cm²).

9. Assuming that the acceleration from *A* to *B* in Fig. 20C is uniform, prove that

$$E_r = -\frac{qa}{c^2 r} \sin \theta$$

where *a* is the acceleration, θ the angle between the direction of motion and the direction of observation, *q* the electric charge, and *r* the distance from the source.

10. Make a polar plot of the intensity radiated from an accelerated charge, assuming the angular dependence of the amplitude to be that given in Prob. 9, which is good for velocities small compared with that of light.

11. Show from the expression for the energy density in an electromagnetic wave that the direction and rate of flow of energy are given by the *Poynting vector* $\mathbf{S} = \frac{c}{4\pi} [\mathbf{E} \times \mathbf{H}]$, where the expression in brackets represents the *vector product*.

CHAPTER 21

SOURCES OF LIGHT AND THEIR SPECTRA

Since light is an electromagnetic radiation, we should expect that the emission of light from any source results from the acceleration of electric charges. It is now certain that the electric charges involved in the emission of visible and ultraviolet light are the negative electrons in the outer part of the atom. By assuming that vibratory or orbital motions of these electrons cause radiation, many of the characteristics of different light sources may be explained. It should be emphasized, however, that this concept must not be carried too far. In the interpretation of spectra it fails in several important respects. These all involve the discrete or corpuscular nature of light, which is to be discussed later (Chap. 30). For the present, we shall confine ourselves to those aspects which can be explained by the assumption that light consists of electromagnetic waves.

21.1. Classification of Sources. Sources of light which are important for optical and spectroscopic experiments may be divided into two main classes: (1) *thermal* sources, in which the radiation is the result of high temperature, and (2) sources depending on the electrical *discharge through gases*. The sun, with its surface temperature of 5000 to 6000°C, is an important example of the first class, but here must also be included such important sources as tungsten-filament lamps, the various electric arcs at atmospheric pressure, and the flame. Under the second class come high-voltage sparks, the glow discharge in vacuum tubes at low pressure, and certain low-pressure arcs like the mercury arc. The distinction between the two classes is not sharp, and we can go continuously from one to the other, for instance by pumping away the air around an electric arc.

21.2. Solids at High Temperature. Almost all practical sources for illuminating purposes use the radiation from a hot solid. In the *tungsten lamp*, the filament is heated to about 2100°C by the dissipation of electrical energy due to its resistance. The filament can be run at temperatures as high as 2300°C but will last for only a short period owing to the rapid vaporization of tungsten. In the *carbon arc* in air, the temperature of the positive pole is about 4000°C and that of the negative pole, 3000°. The positive pole vaporizes and burns away rather rapidly,

but it constitutes the brightest thermal source of light available in the laboratory. The heating results chiefly from the bombardment of the positive pole by electrons drawn from the gaseous part of the arc. Relatively little light comes from the gas itself. An interesting type of arc, useful when a very small source of light is needed, is the so-called *concentrated-arc lamp*. A simplified diagram of this device is shown in Fig. 21A(a). The cathode consists of a small metal tube packed with zirconium oxide, and the anode consists of a metal plate containing a hole slightly larger than the end of the cathode. Tungsten, tantalum, or molybdenum, because of their high melting points, are used for the metal parts. These are sealed into a glass bulb which is filled with an inert gas like argon to a pressure of nearly one atmosphere. The arc runs between the (fused) surface of the zirconium oxide and the surrounding anode, as indicated in part (b) of the figure. The tip of the cathode is heated by ion bombardment to 2700°C or higher, giving it a

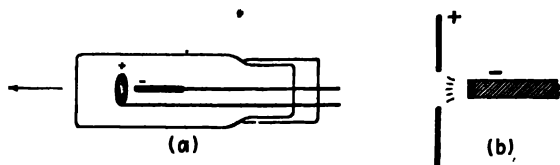


FIG. 21A. The concentrated arc, a close approximation to a "point source."

surface brightness almost equal to that in the carbon arc. The light is observed through the hole in the anode, in the direction shown by the arrow in Fig. 21A(a). Lamps of this type can be made in which the source is as small as 0.007 cm in diameter. A cheaper way of achieving a source of small dimensions is to use a tungsten lamp with a small spiral filament (automobile headlight bulb), run at a voltage somewhat higher than its rated value. This source does not, however, have the smallness and brightness of the concentrated-arc lamp. Other sources of continuous spectra will be considered in Sec. 21.9.

21.3. Metallic Arcs. When two metal rods connected to a source of direct current are touched together and drawn apart, a brilliant arc forms between them. A resistance of high current capacity must be connected in series with the circuit, and adjusted so that the steady current through the arc is from 3 to 5 amp. Higher currents than this will cause excessive heating and melting of the electrodes. A large self-inductance in the circuit will stabilize the arc, and a voltage of 220 is preferable to 110 in this respect. The two poles are held vertically, in line with each other, by clamps with a screw adjustment to vary their separation. In the *iron arc*, the positive pole should be the lower, since

then a bead of molten iron oxide collects in the small cavity which soon forms, and this helps the steadiness of the arc. The radiation from an iron, copper, or aluminum arc comes mostly from the gas traversed by the arc, this gas consisting almost entirely of the vapor of the metal. It has been shown that the gas is at a temperature of from 4000 to 7000°C, and it may in cases of very high currents run up to 12,000°C. The equivalent of a metallic arc may be obtained with a carbon arc in which the positive pole has been bored with an axial hole and packed with the salt of a metal, such as calcium fluoride. It is sometimes desirable to run a metallic arc in an atmosphere other than air by enclosing it in an airtight chamber. The arc may then be run at low pressures as well, but this is a difficult procedure.

Two of the most important types of metallic arcs for laboratory use are the *mercury arc* and the recently developed *sodium arc*. In a common form of mercury arc, liquid mercury is sealed in a highly evacuated glass container of such a shape that the mercury forms two separate pools. These make electrical connection with two wires sealed through the glass. To start the arc, it is tipped until a thread of mercury connects the two pools for an instant and breaks again. As the arc warms up, the pressure of the mercury vapor increases, and unless a fairly large space is available for cooling and condensation, the arc will go out. With sufficient self-inductance in the circuit, the arc may be run at fairly high temperature and pressure, giving a very intense source. For this purpose the container is made of fused quartz to withstand the higher temperature. Quartz has the advantage that it transmits the ultra-violet light (Sec. 22.3), and quartz arcs are frequently used in spectroscopy and for therapeutic purposes. In using them, great care should be taken not to look at the arc too frequently unless glasses are being worn, as a painful inflammation of the eyes may result. The same is true for the exposed metallic arcs mentioned above.

In Fig. 21*B* are shown three types of mercury arc which are convenient for different purposes. That shown in (*a*) is a self-starting type in which the arc is formed in a narrow vertical capillary tube (inside diameter 2 mm) and hence provides an intense vertical line source of mercury light. When connected to the 110-volt d-c mains, the current is limited to about 1.5 amp by the resistances R_1 and R_2 of 80 and 7 ohms, respectively. R_2 is wound on the lower part of the capillary and encased in cement so that it heats the mercury at that point until a bubble of vapor is formed and the mercury thread breaks. The arc immediately starts, generating enough pressure to push the mercury above it up to the point *A*. The arc is then confined to the capillary from *A* to R_2 . The current then falls to about 1.0 amp, owing to the additional resistance

of the arc itself. Although this type is suitable for experiments with small gratings and prism spectroscopes, the pressure and current density in the capillary are high enough to broaden the mercury lines appreciably. Thus, when a source of very monochromatic light* is required, as in work with the Fabry-Perot interferometer, type (a) is not satisfactory. For such purposes type (b) may be used. Here the pool of mercury which constitutes the cathode (negative electrode) may be cooled in water, and the pressure and current density are very low, giving sharp

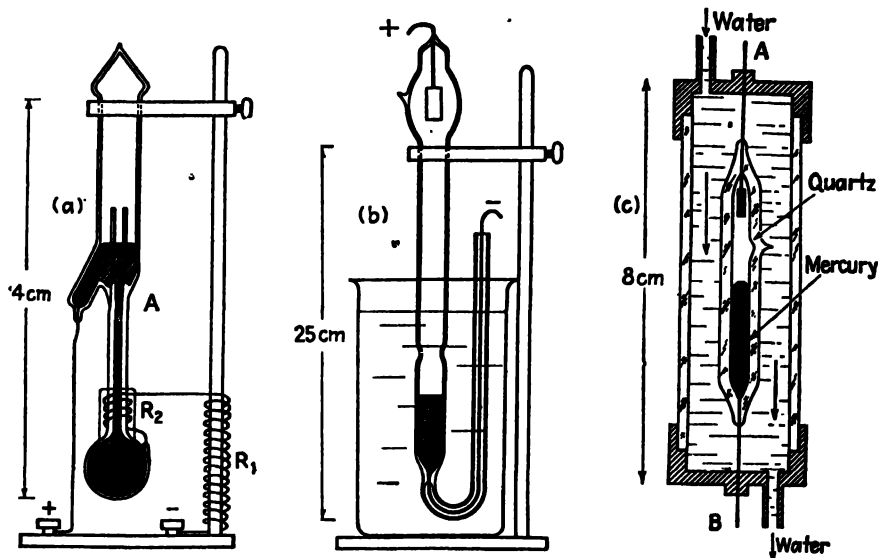


FIG. 21B. Three types of mercury arc. (a) A self-starting mercury arc. (b) A water-cooled arc for spectroscopic research. (c) A high-pressure, high-intensity water-cooled arc for illumination purposes.

lines. The anode is a small iron cylinder. The arc is started by tipping or by holding a wire from a small induction coil next to the glass and at the same time short-circuiting for an instant the inductance which stabilizes the arc once it is established.

In recent years small mercury arcs only 3 or 4 cm long, capable of producing a quarter million candle power, have been made [Fig. 21B(c)]. Such a lamp, constructed originally by Cornelius Bol, consists of a fine quartz capillary which has about 1 mm inside diameter and is half filled with mercury. When a large current is sent through the tube, the

* The green line $\lambda 5461$ of the mercury spectrum is the best for visual work. Its light can be obtained free from the yellow and blue lines by a combination of two glass filters, one of didymium glass (absorbing the yellow lines) and one of bichromate or other yellow glass (absorbing the blue and violet lines).

mercury is quickly vaporized. With so much mercury confined to such a small volume, the vapor pressure, as well as the temperature, becomes extremely high. To prevent the quartz from overheating and melting, water is circulated around the tube as shown. The strong ultraviolet light emitted by the mercury atoms is absorbed by the vapor and reemitted in the visible spectrum as fluorescent radiation. As a consequence, these sources are strong in the visible and relatively weak in the ultraviolet spectrum.

The sodium arc is useful for some purposes, though it is not so intense as the mercury arc. Another disadvantage it has as a source of monochromatic light is due to the double character of the sodium D line. This arc is enclosed in a special type of glass not attacked by sodium vapor. Heating and vaporization of the sodium are brought about by a preliminary glow discharge in one of the rare gases such as argon.

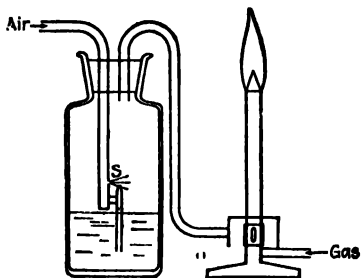


FIG. 21C. Experimental arrangement for producing spectra by introducing salts of metals into the flame of a bunsen burner.

21.4. Bunsen Flame. When sufficient air is admitted at the base of a bunsen burner, the flame is practically colorless, except for a bluish-green cone bounding the inner dark cone of unburnt gas. The temperature above the cone is in the neighborhood of 1800°C , high enough to cause the emission of light from the salts of certain metals when they

are introduced into the flame. The color of the flame and its spectrum are characteristic of the metal and do not depend on which salt is used. The chloride is usually most volatile and gives the most intense coloration. The color of the sodium flame is yellow; of strontium, red; of thallium, green; etc. For introducing the salt into the flame, a common method is to use a loop on the end of a platinum wire, which is first dipped in hydrochloric acid and heated until the sodium yellow disappears. Then, while red-hot, it is touched to the powdered salt, melting a small amount which adheres to the wire. When this is again held in the flame, the color is strong but lasts only a short time. A better method is to mix a fine spray of the chloride solution with the gas before it enters the burner. This is best done with the apparatus shown in Fig. 21C, in case air under pressure is available. Air is forced through the atomizer S, filling the bottle with a fine spray which is carried into the gas at the base of the burner. This gives a very constant light source, and is convenient for the laboratory study of flame spectra. Unfortunately, it can be used for only a limited number of metals, the suitable ones including

lithium, sodium, potassium, rubidium, caesium, magnesium, calcium, strontium, barium, zinc, cadmium, indium, and thallium. Other elements may be used in the hotter oxygen flame or oxyhydrogen flame, but these flames are not so convenient to operate.

21.5. Spark. By connecting a pair of metal electrodes to the secondary of an induction coil or high-voltage transformer, a series of sparks can be made to jump an air gap of several millimeters. If there is no capacity in the circuit, the spark is quiet and not very intense, the radiation coming chiefly from the air in the gap. The spark may be made much more violent and more intense by connecting a condenser (such as a Leyden jar) in parallel across the gap. We then obtain a *condensed spark*. This is an extremely bright source, the spectrum of

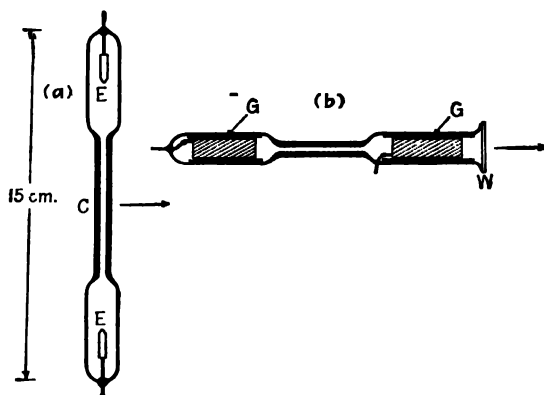


FIG. 21D. Discharge tubes for obtaining the spectra of gases at low pressure.

which is very rich in lines characteristic of the metal of the electrodes. Many of these are so-called “high-temperature” lines, since they are emitted only by a source at effectively high temperature.

21.6. Vacuum Tube. This is a source that has become increasingly common, owing to its application to advertising signs. The familiar red “neon signs” contain pure neon gas at a pressure of about 2 cm Hg. Metal electrodes are sealed through the ends of the tube, and an electric current is caused to traverse the gas by connecting the electrodes to a transformer giving a potential of 5000 to 15,000 volts. Other colors are produced by introducing a small amount of mercury into a neon or argon tube. The heat of the discharge vaporizes the mercury, and we obtain the characteristic color and spectrum of mercury vapor. If the tube is made of colored glass, certain colors of the mercury light are absorbed and various shades of blue and green may be produced.

In the laboratory, this principle can be used on a smaller scale to excite the characteristic radiations of any gas or vapor. Two common forms of vacuum tube are illustrated in Fig. 21D. Type (a) is useful where maximum intensity is not required, for instance if the tube is to be operated with a small induction coil. The electrodes E , E are short pieces of aluminum rod, welded to the ends of tungsten wires, the latter being sealed through the glass. The light is most intense in the capillary tube C , where the current density is greatest, and it is observed laterally, in the direction indicated by the arrow. Considerably greater intensity can be obtained with the "end-on" type shown in (b). Here the electrodes are of sheet aluminum, rolled up and slipped inside two loosely fitting inner glass tubes, G , G . They are fastened to the tungsten leads by wrapping a small strip of aluminum at one end around the wire and pinching it on tightly. The larger area of the electrodes permits the use of greater currents, usually furnished by a transformer, without overheating of the electrodes. The light is observed through a plane glass window W , which may be fused directly to the tube. The inner glass tubes serve to prevent the deposition of aluminum on the outer walls of the main tube, which occurs rather rapidly when a tube is used at a low pressure.*

The exact pressure at which a vacuum tube should be sealed off varies between about 0.5 and 10 mm Hg, according to the gas and to the particular spectrum desired. Only a limited number of gases are suitable for long-continued use in a sealed tube of the above type. Of these, the rare gases neon, helium, and argon are the most satisfactory. Hydrogen, nitrogen, and carbon dioxide tubes will last only a limited time; the gas gradually disappears from the tube, or "cleans up," until a discharge can no longer be maintained.

Starting with a tube filled with some gas at atmospheric pressure, an interesting sequence of phenomena is observed as the pressure is decreased by pumping while the discharge is running continuously. At first a threadlike discharge is obtained very similar to the ordinary spark in air. This gradually broadens out until, when the pressure is 2 or 3 cm, it fills the whole tube. At a few millimeters, *striations* frequently appear throughout the whole discharge, and these separate more and more as

the pressure is lowered still further. Then the main part of the discharge, which contains the striations and is known as the *positive column*, begins to separate from a small bright glow that surrounds the negative electrode and is called the *negative glow*. The *Faraday* dark space* is the region between the negative glow and the first striation of the positive column. At about 1 mm pressure, a small dark region begins to be visible immediately surrounding the cathode. This is the *Crookes† dark space*, and its width furnishes a fairly reliable way of estimating low pressures. This stage of the discharge is shown schematically in Fig. 21E. In an air discharge, the Crookes dark space is 1.2 mm wide for a pressure of 2.06 mm Hg, 2.4 mm wide for 0.6 mm pressure, and 7.0 mm wide for 0.2 mm pressure. When this dark space is fairly wide, a bright sheath of light will be seen covering the surface of the cathode, sometimes called the *cathode glow*. At a pressure low enough so that

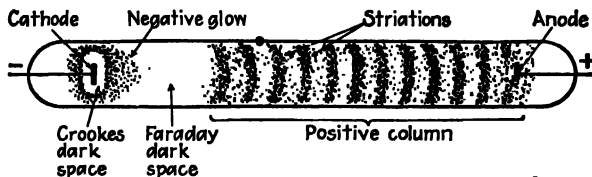


FIG. 21E. General appearance of a high-voltage glow discharge in a gas-filled tube at a pressure of about 1 mm of mercury.

the Crookes dark space touches the glass walls of the tube, the walls begin to fluoresce with a greenish light.

A thorough discussion of the causes of the above complex phenomena, some of which are still imperfectly understood, would take us too far afield. However, the main features of the processes may be briefly stated as follows: The atoms are caused to emit light by the impact of electrons that originate at the surface of the cathode. These are accelerated to high velocities as they traverse the Crookes dark space, because practically the entire potential drop between the electrodes is concentrated across this space. The light of the negative glow is thus produced by very high-speed electrons striking the gas molecules. In the positive column the potential gradient and electron velocities are much smaller. At very low pressures the electrons strike the walls without having

encountered any gas atoms, and their impact causes the fluorescence. Many interesting experiments can be done to show that its origin lies in the fast electrons which fly in straight lines from the cathode and which are called "cathode rays."

21.7. Classification of Spectra. There are two principal classes of spectra, known as *emission spectra* and *absorption spectra*. In each of these there are three types, *continuous*, *line*, and *band spectra*. Emission spectra are obtained when the light coming directly from a source is examined with a spectroscope. Absorption spectra are obtained when the light from a source showing a continuous emission spectrum is passed through an absorbing material and thence into the spectroscope. Figures 21H, 21I, and 21K show reproductions of photographed spectra illustrating the three types, both in emission and in absorption. Solids and liquids, with a few rare exceptions,* give only continuous emission and absorption spectra, in which a wide range of wavelengths, without any sharp discontinuities, is covered. Discontinuous spectra (line and band) are obtained with gases. Gases may also, in certain cases, emit or absorb a true continuous spectrum (Sec. 21.9). The three types of emission spectra may be easily observed with a carbon arc. If the spectroscope is pointed at the white-hot pole of the arc, the spectrum is perfectly continuous. If it is pointed at the violet discharge in the gas between the poles, bands in the green and violet are seen, and there are always a few lines, like the sodium lines, owing to impurities in the carbons.

21.8. Emittance and Absorptance. Although in this chapter we are primarily concerned with various sources of light, and hence with emission, it will be well to state here a very important relation which exists between the emissive and absorptive powers of any surface. A solid, when heated, gives a continuous emission spectrum. The amount of radiation in this spectrum and its distribution in different wavelengths are governed by *Kirchhoff's† law* of radiation. This states that the ratio of the radiant emittance to the absorptance is the same for all bodies at a given temperature. As an equation, this law may be written

$$\frac{W}{a} = \text{const.} = W_b \quad (21a)$$

* Compounds of some of the rare earth metals give line spectra superposed on a continuous spectrum when heated to high temperatures. Their absorption spectra—for example, that of didymium glass—show very narrow regions of absorption, which at liquid-air temperature become sharp absorption lines.

† Gustav Kirchhoff (1824–1887). Professor of physics at Heidelberg and Berlin. Beside discovering some fundamental laws of electricity, he founded (with Bunsen) the science of chemical analysis by spectra.

The quantity W is the total energy radiated per square centimeter of surface per second, while a represents the fraction of the incident radiation which is not reflected or transmitted by the surface. For the constant representing this ratio, we have used the symbol W_b , because it represents the emittance of a so-called *black body*. This term specifies a body which is perfectly black, *i.e.*, one which absorbs all the radiation falling on its surface. Hence for such an ideal body, $a_b = 1$, and W_b equals the constant ratio W/a for other bodies.

Kirchhoff's law expresses a very general relation between the emission and absorption of radiation by the surfaces of different bodies. If the

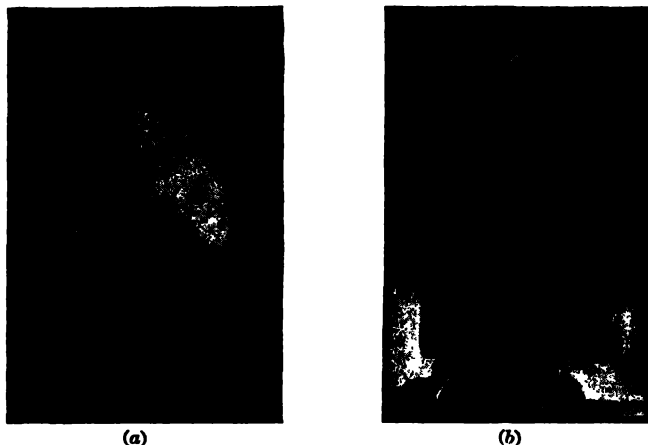


FIG. 21F. Photographs of an electric iron illustrating Kirchhoff's law of radiation. (a) Taken with infrared-sensitive plates with the iron hot but emitting no visible radiation. (b) Taken with ordinary plates and illumination, with the iron at room temperature. (After H. D. Babcock.)

absorptance is high, the emittance must also be high. Here it is essential to realize the difference between the term *absorptance*, which measures the amount of light disappearance at a single reflection, and the *absorption* within the body of the material, as measured by the absorption coefficient α (Sec. 11.7). The latter determines the loss of light upon transmission through the material and has no simple connection with the absorptance of the surface. In the case of metals, for example, we shall see (Sec. 28.14) that a very high absorption coefficient is correlated with a high reflectance. But a high reflectance also means a low absorptance. Thus for metals, and in general for smooth surfaces of pure substances, a high absorption coefficient α necessarily means a low absorptance a .

A black body, which is approximated, for example, by a piece of

carbon, gives the greatest amount of radiation at a given temperature. Transparent or highly reflecting substances are very poor emitters of visible light, even when raised to high temperatures. Figure 21*F* shows a practical illustration of the working of Kirchhoff's law. The right-hand picture is a photograph of an ordinary electric iron at room temperature. A few spots of india ink have been made on the surface, and these appear dark since they are regions of high absorptance. The rest of the surface is highly reflecting and hence a poor absorber. The left-hand photograph was taken by the radiation emitted from the iron when heated. The temperature was less than 400°C, so that no visible radiation was emitted. However, with infrared-sensitive photographic plates a successful photograph was obtained, even though the iron was invisible to the eye in the dark. In this picture, it will be seen that the spots which were previously dark (good absorbers) have now become brighter than the surroundings, even though they have the same temperature. Hence they also emit radiation most copiously, as Kirchhoff's law requires. Here we are assuming that the ink spots, because they are black by visible light, are also good absorbers for infrared light. It is in fact essential that W and a refer to the same wavelength, or range of wavelengths. For the radiation within a small wavelength interval we may write •

$$\frac{W_{\lambda}}{a_{\lambda}} = W_{B\lambda} \quad (21b)$$

indicating by the subscript the emittance and absorptance at a particular wavelength. This form has important applications to discontinuous spectra (Sec. 21.10).

21.9. Continuous Spectra. The most common sources of continuous emission spectra are solids at high temperature, and some of these sources were described in Sec. 21.2. Nothing was said there concerning the distribution in different wavelengths of the energy in the continuous spectrum. According to Kirchhoff's law, this depends on the ability of the surface to *absorb* light of different wavelengths. Thus in a piece of china with a red design glazed upon it, the red parts absorb blue and violet light more strongly than red. When the piece is heated to a high temperature in a furnace and withdrawn, it will be observed that the design will appear bluish by the emitted light, since these portions are the best absorbers and emitters for blue. In general, therefore, the reflectance spectrum of such a solid gives a clue to its emission spectrum.

A black body, which absorbs all wavelengths completely, is commonly taken as the standard because it constitutes a particularly simple case

with which the radiation from other substances may be compared. Figure 21G shows the energy distribution in the radiation from a black body at seven different temperatures, and Fig. 21H(a) shows photographs of the actual spectra corresponding to these curves.* The curve for 2000°K represents fairly well that for a tungsten filament, while that for 6000°K is closely that of the sun (neglecting the narrow regions of absorption due to the Fraunhofer lines). The area under the curve represents the total energy emitted in all wavelengths, and increases

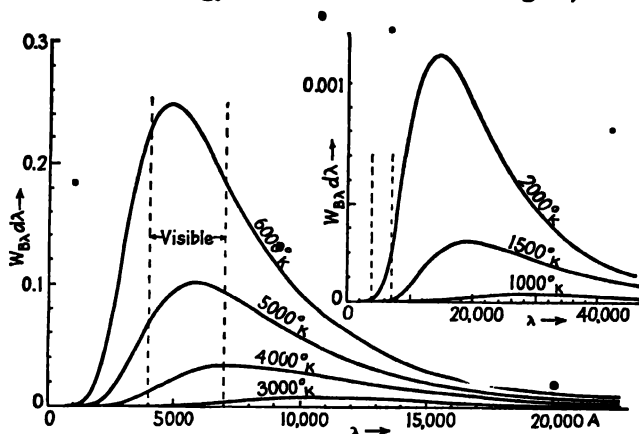


FIG. 21G. Black-body radiation curves (plotted to scale). Abscissas give the wavelengths in angstroms and ordinates the energy in calories per square centimeter per second in a wavelength interval $d\lambda$ of 1 Å. For numerical values, see "Smithsonian Physical Tables," 8th ed., p. 314.

rapidly with the absolute temperature. Calling W_B the total energy in ergs emitted from the surface of a black body per square centimeter per second, and T the absolute temperature, the *Stefan-Boltzmann† law* states that

$$W_B = \sigma T^4 \quad (21c)$$

* In comparing the spectra of Fig. 21H(a) with the curves of Fig. 21G it should be borne in mind that photographed spectra do not reproduce the true distribution of intensity in different wavelengths for three reasons: (1) The dispersion of the prism compresses the spectrum at the long-wavelength end. (2) The photographic plate is not equally sensitive to all wavelengths. In particular, the plate used here does not respond at all beyond $\lambda 6600$. (3) The blackening of the plate is not proportional to the intensity.

† Ludwig Boltzmann (1844–1906). From 1895 to his death by suicide in 1906, professor of physics at Vienna. The law was originally stated by Josef Stefan (1835–1893) and was independently demonstrated theoretically by Boltzmann. The latter is chiefly known for his contributions to the kinetic theory and the second law of thermodynamics.

The constant σ has the value 5.673×10^{-5} erg cm⁻² sec⁻¹°K⁻⁴. The wavelength of the maximum of each curve, λ_{\max} , depends on the temperature according to *Wien's* displacement law*, which states that

$$\lambda_{\max} T = \text{const.} = 0.2897 \text{ cm-deg} \quad (21d)$$

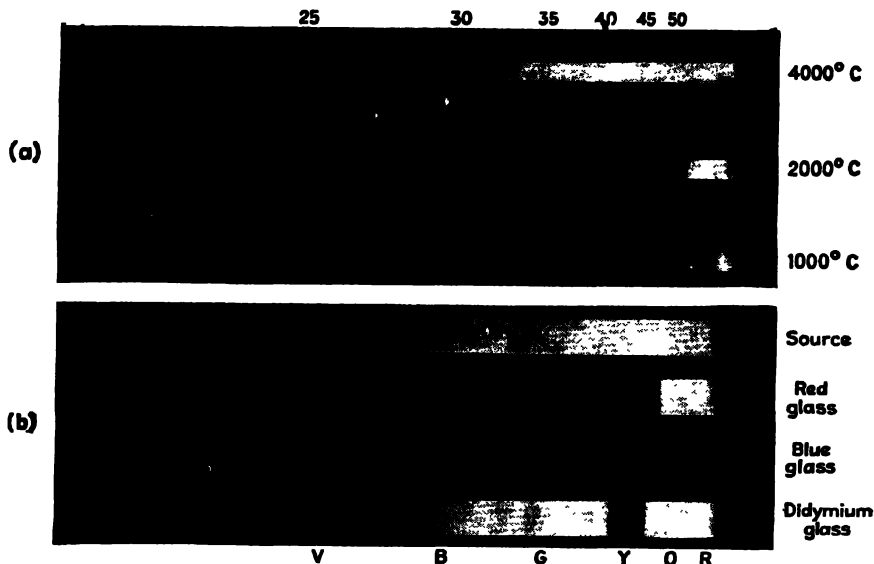


FIG. 21H. Continuous spectra. (a) Continuous emission spectra of a solid at the three temperatures indicated, taken with a quartz spectrograph. The spectra for 1000°C and 2000°C were obtained from a tungsten filament. That for 4000°C is from the positive pole of a carbon arc. The wavelength scale is marked in hundreds of angstroms. (b) Continuous absorption spectra. The upper spectrum is that of the source alone, extending roughly from 4000 to 6500 Å. The others show the effect on this spectrum of interposing three kinds of colored glass.

where λ_{\max} is in centimeters. The shape of the curve itself is given by *Planck's† law*, which may be written

$$W_{B\lambda} d\lambda = \frac{c_1}{\lambda^5} (e^{c_2/\lambda T} - 1)^{-1} d\lambda \quad (21e)$$

Here e is the base of natural logarithms 2.718, while c_1 and c_2 are constants whose values depend on the unit of λ . For λ in centimeters,

$c_1 = 3.7402 \times 10^{-5}$ erg cm² sec⁻¹ and $c_2 = 1.43848$ cm-deg. These constants are of course connected with those in the Stefan-Boltzmann and Wien laws, because Eq. 21c can be obtained from Eq. 21e by integrating it from $\lambda = 0$ to $\lambda = \infty$, while Eq. 21d is obtained if we differentiate Eq. 21e with respect to λ and equate to zero to obtain the maximum value. Thus, the constant in Eq. 21d is $c_2/4.965$. These equations apply, of course, only to the radiation from an *ideal* black body. This can never be strictly realized experimentally, but it is approximated by a black surface or a hollow cavity with a small opening. The quantity $W_{\text{rad}} d\lambda$ denotes the emission of unpolarized radiation per square centimeter per second in all directions in a range $d\lambda$.

A source of a continuous spectrum in the ultraviolet region is sometimes desired for the study of absorption spectra in this region. Hot solids are unsuitable for this purpose, because of the relatively small amount of ultraviolet light they emit, even at the highest temperatures available. It has been found that for this purpose a vacuum-tube discharge through hydrogen gas at 5 to 10 mm pressure is very satisfactory. If a current of a few tenths of an ampere is passed through a tube with a rather wide capillary (5 mm diameter) at 2000 volts, a very intense continuous spectrum is obtained. The maximum intensity of this continuum lies in the violet, but it extends far down into the ultraviolet, to about 1700 Å.

21.10. Line Spectra. When the slit of a prism or grating spectroscope is illuminated with the light from a mercury arc, several lines of different color are seen in the eyepiece. Photographs of common line spectra are shown in Fig. 21I(a) to (j). Each of these lines is an image of the slit formed by the telescope lens by light of a particular wavelength. The different wavelengths are deviated through different angles by the prism or grating; hence the line images are separated. It is important to realize that line spectra derive their name from the fact that a *slit* is customarily used, whose image constitutes the line. If a point, a disk, or any other form of aperture were used in the collimator, the spectrum lines would become points, disks, etc., as the case may be. Frequently, in photographing the spectra from astronomical sources, the collimator is dispensed with entirely, and a prism or grating placed in front of the telescope lens converts the telescope into a spectroscope. In this case, each "line" in the spectrum has the shape of the source. For example, Fig. 21I(h) shows the spectrum of the sun at the instant preceding a total eclipse, when the usual dark-line absorption spectrum is replaced by an emission spectrum from the gases of the solar atmosphere, giving the so-called "flash spectrum." The chief use of a slit is to produce narrow images, so that the images in different wavelengths do not overlap.

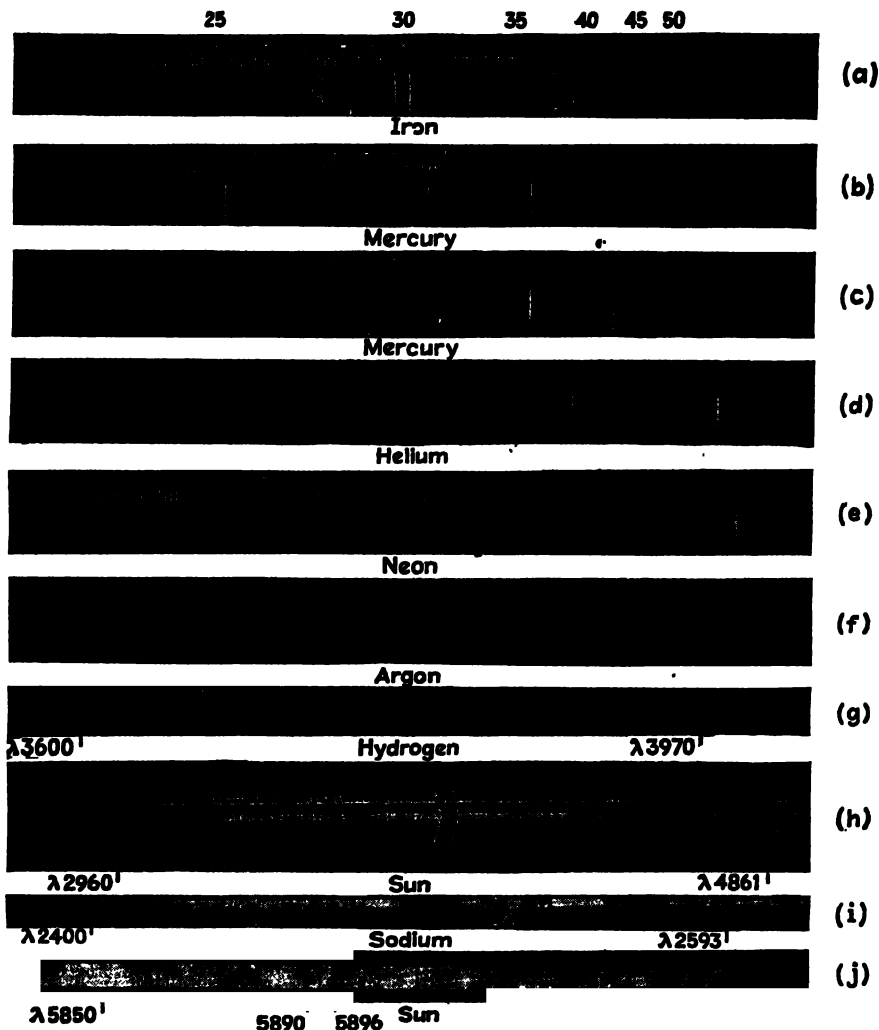


FIG. 217. Line spectra. (a) Spectrum of the iron arc. The line emission spectra (a) to (f) were all taken with the same quartz spectrograph. (b) Mercury spectrum from an arc enclosed in quartz. (c) Same, from an arc enclosed in glass. (d) Helium in a glass discharge tube. (e) Neon in a glass discharge tube. (f) Argon in a glass discharge tube. (g) Balmer series of hydrogen in the ultraviolet, $\lambda\lambda 3600-4000$. This is a grating spectrum. The faint lines on either side of the stronger members are false lines called "ghosts" (Sec. 17.13). (h) Flash spectrum, showing the emission spectrum from the gaseous chromosphere of the sun. This is a grating spectrum taken without a slit at the instant immediately preceding a total eclipse, when the rest of the sun is covered by the moon's disk. The two strongest images are the H and K lines of calcium, and show marked prominences, or clouds of calcium vapor. Other strong images are due to hydrogen and helium. (i) Line absorption spectrum of sodium in the ultraviolet, taken with a grating. The bright lines in the background arise in the source, which here was a carbon arc. Notice the slight continuous absorption beyond the series limit. (j) Solar spectrum in the neighborhood of the D lines. The two strong lines are absorbed by sodium vapor in the chromosphere, and together constitute the first member of the series shown in (i).

The most intense sources of line spectra are metallic arcs and sparks, although vacuum tubes containing hydrogen or one of the rare gases are very suitable. Flames are often used, because the spectra they give are in general simpler, being not so rich in lines. All common sources of line emission or line absorption spectra are gases. Furthermore, it is now known that only the *individual atoms* give true line spectra. That is, when a molecular compound is used in the source, such as methane gas (CH_4) in a discharge tube, or sodium chloride in a "cored" carbon arc, the lines observed are due to the elements and not to the molecules. For example, methane gives a strong line spectrum due to hydrogen, and it is well known that sodium chloride gives the yellow sodium lines. Lines due to carbon and chlorine do not appear with appreciable intensity because these elements are more difficult to excite to emission and their strongest lines lie in the ultraviolet and not in the visible part of the spectrum. In Table 21I are given the wavelengths of the lines in certain commonly used emission spectra, with an indication as to whether they are strong (*s*), medium (*m*), or weak (*w*).

TABLE 21I. WAVELENGTHS, IN ANGSTROM UNITS, OF SOME USEFUL SPECTRAL LINES

Sodium	Mercury	Helium	Cadmium	Hydrogen
5889.95 <i>s</i>	4046.56 <i>m</i>	4387.93 <i>w</i>	4678.16 <i>m</i>	6562.82 <i>s</i>
5895.92 <i>m</i>	4077.81 <i>m</i>	4437.55 <i>w</i>	4799.92 <i>s</i>	4861.33 <i>m</i>
	4358.35 <i>s</i>	4471.48 <i>s</i>	5085.82 <i>s</i>	4340.46 <i>w</i>
	4916.04 <i>w</i>	4713.14 <i>m</i>	6438.47 <i>s</i>	4101.74 <i>w</i>
	5460.74 <i>s</i>	4921.93 <i>m</i>		
	5769.59 <i>s</i>	5015.67 <i>s</i>		
	5790.65 <i>s</i>	5047.74 <i>w</i>		
		5875.62 <i>s</i>		
		6678.15 <i>m</i>		

Line *absorption* spectra are obtained only with gases ordinarily composed of individual atoms (monatomic gases). The absorption lines in the solar spectrum are due to atoms which exist as such, rather than combined as molecules, only because of the high temperature and low pressure in the "reversing layer" of the sun's atmosphere [Fig. 21I(h) and (j)]. In the early days of the study of these lines by Fraunhofer, the more prominent ones were designated by letters. The Fraunhofer lines are very useful "bench marks" in the spectrum, for instance in the measurement and specification of refractive indices. Hence we give here, in Table 21II, their wavelengths and the chemical atoms or molecules to which they are due. The "lines" A, B, and α are really bands, because of absorption by oxygen in the *earth's* atmosphere. It will be

TABLE 21II. THE MOST INTENSE FRAUNHOFER LINES

Designation	Origin	Wavelength, Å	Designation	Origin	Wavelength, Å
A	O ₂	7594-7621 *	b ₄	Mg	5167.343
B	O ₂	6867-6884 *	c	Fe	4957.609
C	H	6562.816	F	H	4861.327
α	O ₂	6276-6287 *	d	Fe	4668.140
D ₁	Na	5895.923	e	Fe	4383.547
D ₂	Na	5889.953	f	H	4340.465
D ₃	He	5875.618	G	Fe	4307.906
F ₂	Fe	5269.541	G	Ca	4307.741
b ₁	Mg	5183.618	g	Ca	4226.728
b ₂	Mg	5172.699	h	H	4101.735
b ₃	Fe	5168.901	H	Ca ⁺	3968.468
b ₄	Fe	5167.491	K	Ca ⁺	* 3933.666

* Band.

seen that b₄ and G are blends of two lines which are not ordinarily resolved but are due to different elements.

In the laboratory, there are only a few substances which are suitable for observing line absorption spectra, because the absorption lines of most monatomic gases lie far in the ultraviolet. The alkali metals are one exception, and if sodium is heated in an evacuated steel or pyrex-glass tube with glass windows at the ends, the spectrum of light from a tungsten source viewed through the tube will show the sodium lines in absorption [Fig. 21I(i)]. They appear as dark lines against the ordinary continuous emission spectrum.

A somewhat simpler experiment to perform, and one which in addition shows the application of Kirchhoff's law to line spectra, is illustrated diagrammatically in Fig. 21J. A is a horizontal carbon arc cored with

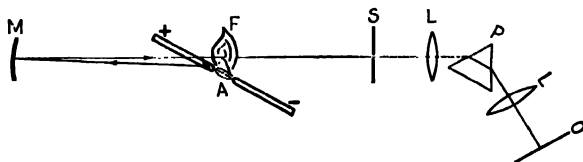


FIG. 21J. Experimental arrangement for showing absorption of the sodium D lines, and for illustrating Kirchhoff's law of radiation.

sodium chloride. The arc is run on a fairly large current so that a bright yellow flame *F* rises above it. If the slit *S* of the spectroscope is directed at the flame, the sodium D lines are seen in emission. They can now

be observed in absorption by placing a concave mirror M in such a position that it casts an image of the bright positive pole of the arc on the slit, the light passing through the flame on its way to the slit. There is a considerable concentration of sodium atoms in the flame, and these are able to absorb, as well as to emit, the particular frequencies corresponding to the D lines. Under these circumstances the lines appear dark in the spectrum, because of the fact that the flame is *at a lower temperature* than the positive pole. This is a consequence of Kirchhoff's law in the form of Eq. 21b. To show this, suppose the absorptance a_λ of the flame for the wavelength of the D lines to be $\frac{1}{4}$, so that one-quarter of this radiation coming from the mirror is removed from the beam. But according to Eq. 21b W_λ for this wavelength is $\frac{1}{4}W_{\lambda\lambda}$, that is, the yellow lines are emitted with one-quarter of the intensity of the corresponding portion of the radiation from a black body *at the temperature of the flame*. Hence if the pole of the arc were at the same temperature as the flame, the amount absorbed would be just compensated by the emission, and no line would appear in the spectrum.* The pole, however, is at a considerably higher temperature; hence the amount absorbed is greater than that emitted by the flame, and dark lines are actually observed with the mirror in position. By shifting the mirror so that the image of a cooler part of the pole falls on the slit, the lines can be made to disappear, or to change into bright lines when the temperature of the selected part of the pole is less than that of the flame.

21.11. Theory of the Connection between Emission and Absorption. Kirchhoff's law, as stated in Sec. 21.8, may be proved rigorously by thermodynamical methods. However, it will help more in understanding the above experiment to consider the processes of emission and absorption from the electromagnetic standpoint. We may picture the emission of light as due to periodic motions of the electrons in the atoms of the source. These motions will cause electromagnetic waves to be sent out having the same frequencies as the charged particles, just as the sound emitted from a tuning fork has the frequency of the fork. In the case of sodium vapor, each oscillating charge vibrates with a particular frequency, like the tuning fork, and the frequency is that of the yellow sodium light. Now if we consider sodium light to be sent through the vapor, the analogy with the tuning fork is still valid. It is well known that when sound waves of the right frequency are incident on a tuning fork, the fork will start vibrating and will pick up a considerable amplitude by virtue of resonance. In the same way the sodium atoms

* We are assuming here that the pole radiates as a perfect black body.

respond to the incident electromagnetic waves, and the energy which they absorb from the waves is reemitted as *resonance radiation*. Although all the energy taken from the waves is thus reemitted, resonance radiation is uniformly distributed in all directions and thus will be relatively weaker in the forward direction than if the absorbing atoms were not present.

The connection between the emittance and absorptance of a substance for light of a given wavelength necessarily follows from the above considerations. If a substance absorbs light of one frequency strongly, it must possess a large number of charges whose characteristic frequencies of vibration match that of the light. Conversely, when the substance is caused to emit light, these same vibrations will cause strong emission of the same frequency.

21.12. Series of Spectral Lines. In the spectra of some elements, lines are observed which obviously belong together to form a *series* in which the spacing and intensities of the lines change in a regular manner. For example, in the Balmer series of hydrogen [Fig. 217(*g*)] the spacing of the lines decreases steadily as they proceed into the ultraviolet toward shorter wavelengths, and their intensities fall off rapidly. Although only the first four lines lie in the visible region, the Balmer series has been traced by photography to 31 members in the spectra of hot stars, where it appears as a series of absorption lines. The absorption spectrum of sodium vapor shows a remarkably long series of lines, each of which is a close doublet [not resolved in Fig. 217(*i*)], known as the *principal series*. This series also appears in emission from the arc or flame, and the well-known D lines constitute the first doublet of the series. In the sodium spectrum from a flame, about 97 per cent of the intensity in this series is in the first member. The emission spectra of the alkalis also show two other series of doublets in the visible region, known as the *sharp* and *diffuse* series. A fourth weak series in the infrared is called the *fundamental* series. The alkaline earth metals, such as calcium, show two such sets of series—one of single lines, the other of triplets.

A characteristic of any particular series is the approach of the higher series members to a certain limiting wavelength, known as the *limit* or *convergence* of the series. In approaching this limit, the lines crowd closer and closer together, so that there are theoretically an infinite number of lines before the limit is actually reached. Beyond the limit a rather faint continuous spectrum is sometimes observed in emission; in absorption a region of continuous absorption can always be observed if the absorbing vapor is sufficiently dense [Fig. 217(*i*)]. The series

limits furnish the clue to the identification of the type to which the series belongs. Thus the sharp and diffuse series approach the same limit, while the principal series approaches another limit which for the alkalis lies at shorter wavelengths.

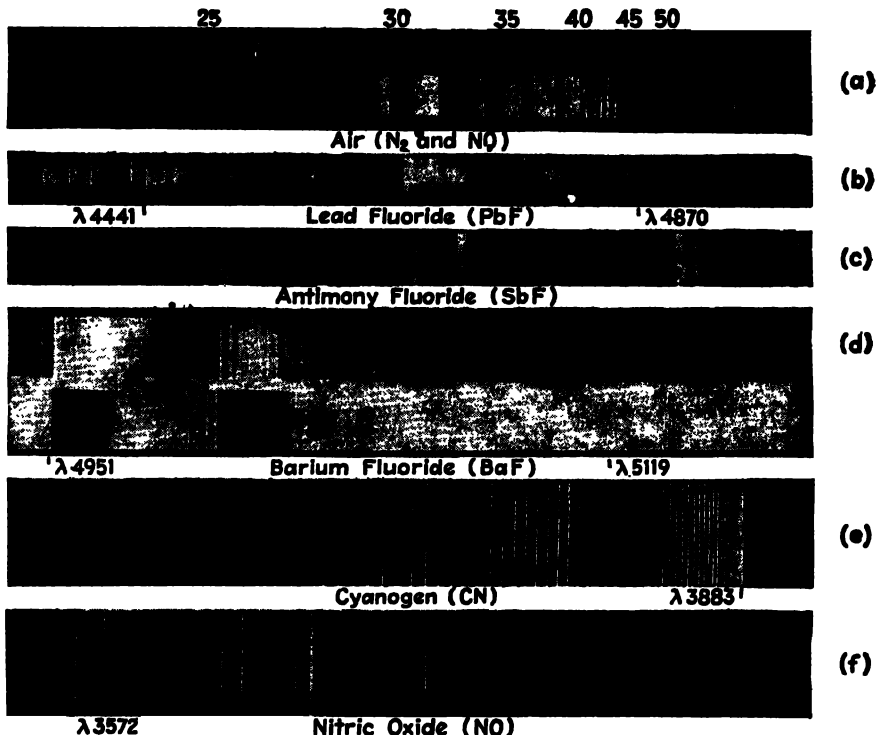


FIG. 21K. Band spectra. (a) Spectrum of a discharge tube containing air at low pressure. Four band systems are present: the γ -bands of NO ($\lambda\lambda 2300$ -2700), negative nitrogen bands (N_2^+ , $\lambda\lambda 2900$ -3500), second positive nitrogen bands (N_2 , $\lambda\lambda 2900$ 5000), and the first positive nitrogen bands (N_2 , $\lambda\lambda 5500$ 7000). (b) Spectrum of a high-frequency discharge in lead fluoride vapor. These bands, due to PbF, fall in prominent sequences. (c) Spectrum showing part of one band system of SbF, obtained by vaporizing antimony fluoride into "active nitrogen." (b) and (c) were taken with a large quartz spectrograph. (d) Emission and absorption band spectra of BaF. Emission from a carbon arc cored with BaF₂; absorption of BaF vapor in an evacuated steel furnace. The bands are closely grouped in sequences. Second order of 21-ft. grating. (e) CN band at $\lambda 3883$ from an argon discharge tube containing carbon and nitrogen impurities. Second order of grating. (f) Band in the ultraviolet spectrum of NO, obtained from glowing "active nitrogen" containing a small amount of oxygen. Second order of grating. [(b) and (c) after Rochester.]

21.13. Band Spectra. The most convenient sources of band spectra for laboratory observation are the carbon arc cored with a metallic salt, the vacuum tube, and the flame. Calcium or barium salts are suitable in the arc or flame, and carbon dioxide or nitrogen in a vacuum tube.

As observed with a spectroscope of small dispersion, these spectra present a typical appearance which distinguishes them at once from line spectra [Fig. 21K(a) to (d)]. Many bands are usually observed, each with a sharp edge on one side called the *head*. From the head, the band shades off gradually on the other side. In some band spectra several closely adjacent bands, overlapping to form *sequences*, will be seen [Fig. 21K(b) and (d)], while in others the bands are spaced fairly widely, as in Fig. 21K(c). When the high dispersion and resolving power of a large grating are used, each band is found to be actually composed of many fine lines, arranged with obvious regularity into series called *branches* of the band. In Fig. 21K(e), two branches will be seen starting in opposite directions from a pronounced gap, where no line appears. In (f) the band is double, and the two branches of the left-hand member can be seen running side by side.

Various sorts of evidence point to the conclusion that band spectra arise from *molecules*, i.e., combinations of two or more atoms. Thus it is found that, while the atomic or line spectrum of calcium is independent of which salt we put in the arc, we obtain different bands by using calcium fluoride, calcium chloride, or calcium bromide. Also, the bands appear in those types of sources where the gas receives less violent treatment. Nitrogen in a vacuum tube subjected to an ordinary uncondensed discharge shows only the band spectrum, whereas if a condensed discharge is used, the line spectrum appears. The most conclusive evidence lies in the fact that the absorption spectrum of a gas which is known to be molecular (O_2 , N_2) shows bands but no lines, owing to the absence of any dissociation into atoms. Furthermore, it is found that any simple band spectrum, like those described and illustrated above, is due to a *diatomic* molecule. When calcium fluoride (CaF_2) is put into the arc, the bands observed are due to CaF . The violet bands in the uncored carbon arc are due to CN , the nitrogen coming from the air [Fig. 21K(e)]. Carbon dioxide in a vacuum tube gives the spectrum of CO , and there are many other examples of this type of dissociation of the more complex molecules into diatomic ones.

21.14. Theory of Line, Band, and Continuous Spectra. The attempt to interpret the various definite frequencies emitted by the atoms of a gas in producing a line spectrum has constituted what is probably the most remarkable chapter in the history of physics. Just as the frequencies of vibration of a violin string give sound waves whose frequencies bear the simple ratio of whole numbers to the fundamental note, it was first supposed that the frequencies of the light in the various spectral lines should bear some definite relation to each other, which would furnish the clue to the modes of vibration of the atom and to its structure.

This has proved to be the case, though in a very different way than was at first anticipated. The definite relation of frequencies is actually found in spectral series. However, it will be seen at once that the atomic frequencies do not behave like those of a violin string. In the latter the overtones increase steadily toward an infinite frequency (zero wavelength), while the frequencies in a spectral series approach a definite limiting value. The complete explanation of line spectra has now been obtained by developing an entirely new theory, called the *quantum theory*.* Although this theory is in many respects in direct contradiction to the electromagnetic theory, the latter proved an invaluable guide in attacking such problems as the intensity and polarization of spectral lines. It also gave the first clue to the behavior of the lines when the source was placed in a magnetic field (Chap. 29). For the complete explanation of line spectra, however, the quantum theory is absolutely essential. We shall return to this subject in the final chapter.

Band spectra have also required the quantum theory for their complete explanation. Nevertheless, the electromagnetic treatment of the problem of molecular spectra was successful in some respects. Certain series of bands are observed in the infrared which have frequencies and intensities related very closely like a fundamental and overtones. These are now known to be due to the vibration of the two nuclei in a diatomic molecule along the line joining them. The two branches of an individual band [Fig. 21K(e)] could be explained as due to rotation of the molecule about a direction perpendicular to the above line. Thus the electromagnetic theory predicts two combination frequencies, the sum and the difference of the frequencies of vibration and rotation. This theory, however, required a continuous distribution of frequencies in each branch, and was unable to explain the discrete lines.

That a continuous spectrum is obtained from liquids and solids can be understood from the fact that here the atoms are closer together than in a gas and exert forces on each other. Whereas in a gas the atoms are far apart and able to emit definite frequencies, these are so modified by the mutual influence of the atoms in a solid that they are spread out into a continuous spectrum. The beginning of this effect is seen in the spectrum of a gas at a fairly high pressure. The lines become broadened, owing to the frequent collisions of the atoms, and the broadening increases with pressure, so that finally the lines merge into a continuous spectrum

as the gas approaches the liquid state. On the electromagnetic theory, one can understand qualitatively the increase in the radiation from a solid with increase of temperature. The motions of the charged particles increase in amplitude as the substance becomes hotter, with a resultant increase in amplitude of the emitted waves. More rapid accelerations would cause the average wavelength to shift toward higher frequencies as the temperature is raised. Again, however, the quantum theory is required to explain the actual distribution of energy in different wavelengths. In fact, it was the attempt to derive Eq. 21e which first led Planck to make the revolutionary assumptions which constituted the foundations of this theory. For further discussion of the quantum aspects, see Chap. 30.

Problems

1. Calculate to four significant figures the wavelength λ_{\max} of the maximum for the seven curves shown in Fig. 21G.
2. Calculate to three significant figures the value of the energy maximum for the 5000°K curve shown in Fig. 21G.
3. Find the total energy radiated in 2 min from a cube of copper 3 cm on edge, when the temperature is 1500°F. Assume an absorptivity of 0.92.
4. If the surroundings in Prob. 3 are at a temperature of 100°F, how much energy is absorbed by the copper block in 2 min? Compare this energy with that emitted. (NOTE: If both cube and surroundings were at 100°F, the cube would have to absorb the same energy that it emits.)
5. On mapping the spectral intensity curve of an incandescent source, it is found to have a maximum at a wavelength of 22,000 Å. What is the temperature of the source?
6. A thermopile is placed near a furnace having a small circular opening. When the furnace is at 2050°K, and its distance from the thermopile such that the hole subtends the same solid angle as the sun's disk, the galvanometer deflection is found to be $\frac{1}{16}$ of that when the thermopile is placed in full sunlight. Find the temperature of the sun indicated by this experiment.
7. Which line or lines of the helium spectrum are absorbed by didymium glass? By the red glass whose spectrum is shown in Fig. 21H?
8. By inspection of Fig. 21I(b) and (c), sketch a curve showing how the absorption of glass changes with wavelength through the visible and ultraviolet.
9. It is desired to make neon tubes containing mercury so that the tubes will give (a) a green color, (b) a blue color. What characteristics are desired for the glass of the tube in each case? Specify the ranges of wavelengths absorbed.
10. The discharge tube in Fig. 21D(a), containing air at a low pressure, is steadily exhausted with a pump. At approximately what pressure will the glass walls begin to fluoresce by the impact of cathode rays?
11. A gray surface has an absorptance $a = 0.4$. Find the rate of emission in ergs per square centimeter per second at a temperature of 90°C.
12. Derive Eq. 21d from Eq. 21e, as suggested in Sec. 21.9.
13. Derive Eq. 21c from Eq. 21e, as suggested in Sec. 21.9.
14. In the experiment of Sec. 21.10, the apparatus is adjusted to show a dark sodium line in the bright continuous spectrum. Interposing a piece of gray glass,

which absorbs three-quarters of the light of any wavelength, between the arc and the mirror changes the dark line to a bright line. Explain.

15. If the λ_{\max} radiated from a black body is at 500 Å, what is its absolute temperature? How much energy will the body radiate in 1 sec from each square centimeter of its surface?

16. Solve Prob. 15 if λ_{\max} is at 2000 Å.

17. It has been proposed that the radio "noise" observed in the microwave region at $\lambda = 1$ cm is thermal radiation from the sun. Assuming that the sun radiates as a black body at 6000°K, and that the disk of the sun subtends an angle of $\frac{1}{2}^\circ$, find the absolute intensity in ergs per square centimeter per second to be expected for such radiation.

18. Suppose that the red hydrogen line at 6563 were the fundamental frequency of a small solid vibrator representing the atom. Plot on a wavelength scale the spectrum to be expected for hydrogen if it were to radiate this frequency plus all its harmonics. Compare with Fig. 211(g).

19. An effective demonstration of Kirchhoff's law of radiation is obtained by fusing a little didymium into a thin quartz rod and heating it in a bunsen flame. Describe the emission spectrum that would be expected in this experiment.

CHAPTER 22

ABSORPTION AND SCATTERING

When a beam of light is passed through matter in the solid, liquid, or gaseous state, its propagation is affected in two important ways. In the first place, the intensity will always decrease to a greater or less extent as the light penetrates farther into the medium. In the second place, the velocity will be less in the medium than in free space. The loss of intensity is chiefly due to absorption, although under some circumstances scattering may play an important part. In this chapter we shall discuss the consequences of absorption and scattering, while the effect of the medium on the velocity, which comes under the term "dispersion," we shall consider in the following chapter. The term absorption as used in this chapter refers to the decrease of intensity of light as it passes through a substance (Sec. 11.7). It is important to distinguish this definition from that of absorptance, which was given in Sec. 21.8. The two terms refer to different physical quantities, but there are certain relations between them, as we shall now see.

22.1. General and Selective Absorption. A substance is said to show *general absorption* if it reduces the intensity of all wavelengths of light by nearly the same amount. For visible light this means that the transmitted light, as seen by the eye, shows no marked color. There is merely a reduction of the total intensity of the white light, and such substances therefore appear to be gray. No substance is known which absorbs all wavelengths equally, but some, such as suspensions of lamp black or thin semitransparent films of platinum, approach this condition over a fairly wide range of wavelengths.

By *selective absorption* is meant the absorption of certain wavelengths of light in preference to others. Practically all colored substances owe their color to the existence of selective absorption in some part or parts of the visible spectrum. Thus a piece of green glass absorbs completely the red and blue ends of the spectrum, the remaining portion in the transmitted light giving a resultant sensation of green to the eye. The colors of most natural objects such as paints, flowers, etc., are due to selective absorption. These objects are said to show pigment or *body color*, as distinguished from *surface color*, since their color is produced by light which penetrates a certain distance into the substance. Then,



by scattering or reflection, it is deviated and escapes from the surface, but only after it has traversed a certain thickness of the medium and has been robbed of the colors which are selectively absorbed. In all such cases the absorptance of the body will be proportional to its true absorption and will depend in the same way upon wavelength. Surface color, on the other hand, has its origin in the process of reflection at the surface itself (Sec. 22.7). Some substances, particularly metals like gold or copper, have a higher reflecting power for some colors than for others, and therefore show color by reflected light. The transmitted light here has the complementary color, whereas in body color the color is the same for the transmitted and reflected light. A thin gold foil, for example, looks yellow by reflection and blue green by transmission. As was mentioned in Sec. 21.8, the body absorption of these materials is very high. This causes a high reflectance and a correspondingly low absorptance.

22.2. Distinction between Absorption and Scattering. In Fig. 22A let light of intensity I_0 enter a long glass cylinder filled with smoke.

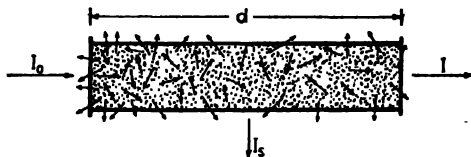


FIG. 22A. Scattering of light by finely divided particles such as those in smoke.

The intensity I of the beam emerging from the other end will be less than I_0 . For a given density of smoke, experiment shows that I depends on the length d of the column according to the exponential law stated in Sec. 11.7, i.e.,

$$I = I_0 e^{-\alpha d} \quad (22a)$$

Here α is usually called the absorption coefficient, since it is a measure of the rate of loss of light from the direct beam. However, most of the decrease of intensity of I is in this case not due to a real disappearance of the light, but results from the fact that some light is *scattered* to one side by the smoke particles and thus removed from the direct beam. Even with a very dilute smoke, a considerable intensity I_s of scattered light may easily be detected by observing the tube from the side in a darkened room. Rays of sunlight seen to cross a room from a window are made visible by the fine suspended dust particles present in the air.

True *absorption* represents the actual disappearance of the light, the energy of which is converted into *heat motion* of the molecules of the absorbing material. This will occur to only a small extent in the above experiment, so that the name "absorption coefficient" for α is not appro-

priate in this case. In general, we can regard α as made up of two parts, α_s due to true absorption, and α_{sc} due to scattering. Equation 22a then becomes

$$I = I_0 e^{-(\alpha_s + \alpha_{sc})d} \quad (22b)$$

In many cases either α_s or α_{sc} may be negligible with respect to the other, but it is important to realize the existence of these two different processes and the fact that in many cases both may be operating.

22.3. Absorption by Solids and Liquids. If monochromatic light is passed through a certain thickness of a solid or of a liquid enclosed in a transparent cell, the intensity of the transmitted light may be much smaller than that of the incident light, owing to absorption. If the wavelength of the incident light is changed, the amount of absorption will also change to a greater or less extent. A simple way of investigating the amount of absorption for a wide range of wavelengths simultaneously is illustrated in Fig. 22B. S_1 is a source which emits a continuous range

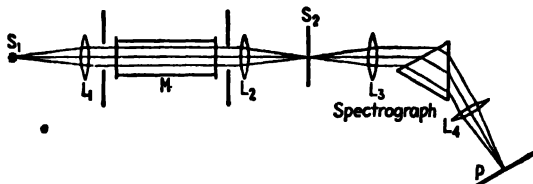


FIG. 22B. Experimental arrangement for observing the absorption of light by solids, liquids, or gases.

of wavelengths, such as an ordinary tungsten-filament lamp. The light from this source is rendered parallel by the lens L_1 and traverses a certain thickness of the absorbing medium M . It is then focused by L_2 on the slit S_2 of a prism spectrograph, and the spectrum is photographed on the plate P . If M is a "transparent" substance like glass or water, the part of the spectrum on P representing visible wavelengths will be perfectly continuous, as if M were not present. If M is colored, part of the spectrum will be blotted out, corresponding to the wavelengths removed by M , and we call this an *absorption band*. For solids and liquids, these bands are almost always continuous in character, fading off gradually at the ends. Examples of such absorption bands were shown in Fig. 21H(b).

Even a substance which is transparent to the visible region will show such selective absorption if the observations are extended far enough into the infrared or the ultraviolet region. Such an extension involves considerable experimental difficulty when a prism spectrograph is used, because the material of the prism and lenses (usually glass) may itself

have strong selective absorption in these regions. Thus flint glass cannot be used much beyond 25,000 Å, (or 2.5 μ) in the infrared, nor beyond about 3800 Å in the ultraviolet. Quartz will transmit somewhat farther in the infrared and much farther in the ultraviolet. Table 22I shows the limits of the regions over which various transparent substances used for prisms will transmit an appreciable amount of light.

TABLE 22I

Substance	Limit of transmission, Å	
	Ultraviolet	Infrared
Crown glass.....	3500	20,000
Flint glass.....	3800	25,000
Quartz (SiO_2).....	1800	40,000
Fluorite (CaF_2).....	1250	95,000
Rock salt (NaCl).....	1750	145,000
Sylvin (KCl).....	1800	230,000
Lithium fluoride.....	1100	70,000

Prisms for infrared investigations are usually of rock salt, while for the ultraviolet quartz is most common. In an ultraviolet spectrograph, there is no advantage in using fluorite unless air is completely removed from the light path, because this begins to absorb strongly below 1850 Å. Also, specially prepared photographic plates must be used below this wavelength, since the gelatin of the emulsion by its absorption renders ordinary plates insensitive. In the infrared, photography can now be used as far as 13,000 Å, thanks to recently developed methods of sensitizing plates. Beyond this, an instrument based upon measurement of the heat produced, such as a thermopile, must be used.

When absorption measurements are extended over the whole electromagnetic spectrum, it is found that no substance exists which does not show strong absorption for some wavelengths. The metals exhibit general absorption, with a very minor dependence on wavelength in most cases. There are exceptions to this, however, as in the case of silver, which has a pronounced "transmission band" near 3160 Å (see Fig. 28M). A film of silver which is opaque to visible light may be almost entirely transparent to ultraviolet light of this wavelength. Dielectric materials, which are poor conductors of electricity, exhibit pronounced selective absorption which is most easily studied when scattering is avoided by having them in a homogeneous condition such as that of a single crystal, a liquid, or an amorphous solid. In a general way, it may be said that such substances are more or less transparent to X rays and

γ rays, *i.e.*, light waves of wavelength below about 10 Å. Proceeding toward longer wavelengths, we encounter a region of very strong absorption in the extreme ultraviolet, which in some cases may extend to the visible region, or beyond, and in others may stop somewhere in the near ultraviolet (see Table 22I). In the infrared, further absorption bands are encountered, but these eventually give way to almost complete transparency in the region of radio waves. Thus for dielectrics we may usually expect three large regions of transparency, one at the shortest wavelengths, one at intermediate wavelengths (perhaps including the visible), and one at very long wavelengths. The limits of these regions vary a great deal in different substances, and one substance, such as water, may be transparent to the visible but opaque to the near infrared, while another, such as rubber, may be opaque to the visible but transparent to the infrared.

22.4. Absorption by Gases. The absorption spectra of all gases at ordinary pressures show narrow absorption lines. In certain cases it is also possible to find regions of continuous absorption (Sec. 21.12), but the outstanding characteristic of gaseous spectra is the presence of these sharp lines. If the gas is monatomic like helium or mercury vapor, the spectrum will be a true line spectrum, frequently showing clearly defined series. The number of lines in the absorption spectrum is invariably less than in the emission spectrum. For instance, in the case of the vapors of the alkali metals, only the lines of the principal series are observed under ordinary circumstances [Fig. 21I(i)]. The absorption spectrum is therefore simpler than the emission spectrum. If the gas consists of diatomic or polyatomic molecules, the sharp lines form the rotational structure of the absorption bands characteristic of molecules. Here again the absorption spectrum is the simpler, and fewer bands are observed in absorption than in emission from the same gas [Fig. 21K(d)].

22.5. Resonance and Fluorescence of Gases. Let us consider what happens to the energy of incident light which has been removed by the gas. If true absorption exists, according to the definition of Sec. 22.2, this energy will all be changed into heat, and the gas will be somewhat warmed. Unless the pressure is very low, this is generally the case. After an atom or molecule has taken up energy from the light beam, it may collide with another particle, and an increase in the average velocity of the particles is brought about in such collisions. The length of time that an energized atom can exist as such before a collision is only about 10^{-7} or 10^{-8} sec, and unless a collision occurs before this time, the atom will get rid of its energy as radiation. At low pressures, where the time between collisions is relatively long, the gas will become a secondary

source of radiation, and we do not have true absorption. The reemitted light in such cases usually has the same wavelength as the incident light, and is then termed resonance radiation (Sec. 21.11). This radiation was discovered and extensively investigated by R. W. Wood.* The origin of its name is clear, since as has been mentioned the phenomenon is analogous to the resonance of a tuning fork. Under some circumstances the reemitted light may have a longer wavelength than the incident light. This effect is called *fluorescence*. In either resonance or fluorescence, some of the light is removed from the direct beam and dark lines will be produced in the spectrum of the transmitted light. Resonance and fluorescence are not to be classed as scattering. This distinction will be made clear in Sec. 22.12.

Resonance radiation from a gas can readily be demonstrated by the use of a sodium-arc lamp. A small lump of metallic sodium is placed in a glass bulb connected to a vacuum pump. The sodium is distilled from one part of the bulb to another by heating with a bunsen burner, thus liberating the large quantities of hydrogen always contained in this metal. After a high vacuum is attained, the bulb is sealed off and the light of the arc is focused by a lens on the bulb. The bulb must of course be observed from the side in a dark room. On gently warming the sodium with the flame, a cone of yellow light defining the path of the incident light will be seen. At higher temperatures, the glowing cone becomes shorter, and eventually is seen merely as a thin bright skin on the inner surface of the glass.

Fluorescence of a gas is most easily shown with iodine vapor, which consists of diatomic molecules, I_2 . White light from a carbon arc will produce a greenish cone of light when focused in a bulb containing iodine vapor in vacuum at room temperature. A still more interesting experiment can be performed by using monochromatic light from a mercury arc, as shown in Fig. 22C. The source of light is a long horizontal arc *A*, which is enclosed in a box with a long slot cut in the top parallel to the arc. Immediately above this is a glass tube *B* filled with water. This acts as a cylindrical lens to concentrate the light along the axis of tube *C*, containing the iodine vapor in vacuum. The fluorescent light from the vapor is observed with a spectroscope pointed at the plane window on the end of tube *C*. The other end is tapered and painted black to prevent reflected light from entering the spectroscopic, and a

* R. W. Wood (1868-). Professor of experimental physics at the Johns Hopkins University. He pioneered in many fields of physical optics and also became one of the most colorful figures in American physics. His discoveries in optics are contained in his excellent text "Physical Optics."

screen with a circular hole placed close to the window helps in this respect. A polished reflector *R* laid over *C* increases the intensity of illumination. If *B* contains a solution of potassium dichromate and neodymium sulfate, only the green line of mercury, $\lambda 5461$, is transmitted. Figure 22D(b)

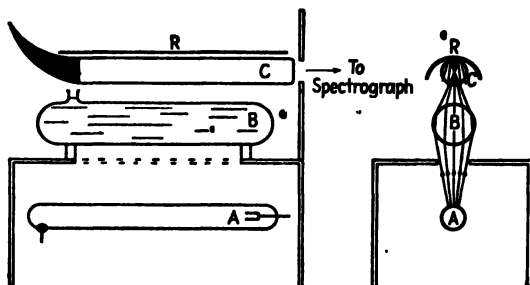


FIG. 22C. Experimental arrangement for observing the fluorescence of iodine vapor with excitation by monochromatic light.

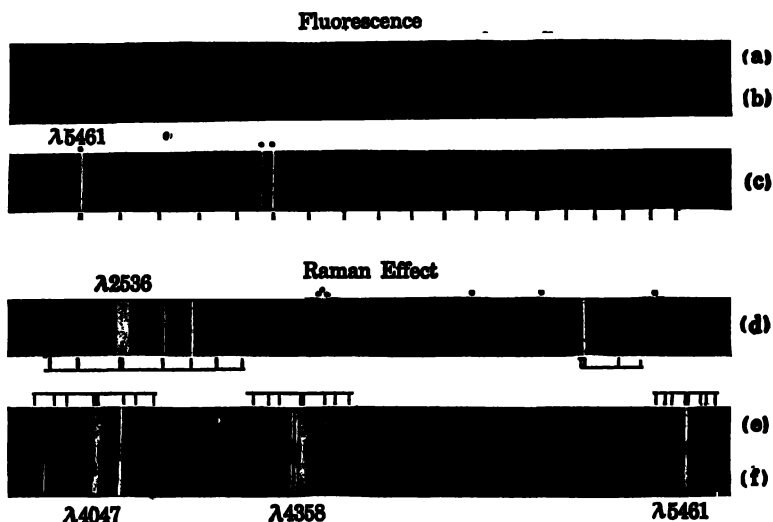


FIG. 22D. Photographs of (a) mercury-arc spectrum; (b) fluorescence spectrum of iodine, (c) enlarged section of (b); (d) Raman spectrum of hydrogen (after Rasetti); (e) Raman spectrum of liquid carbon tetrachloride (after M. Jeppeson); (f) mercury arc.

and (c) were reproduced from a spectrogram taken in this way, though with water in the tube *B*. Beside the lines of the ordinary mercury spectrum (marked by dots in the figure) which are present as a result of ordinary reflection or Rayleigh scattering (Sec 22.10), one observes a series of almost equally spaced lines extending toward the red from the green line. These represent the fluorescent light of modified wavelength.

22.6. Fluorescence of Solids and Liquids. If a solid or a liquid is strongly illuminated by light which it is capable of absorbing, it may reemit fluorescent light. According to *Stokes' law*, the wavelength of the fluorescent light is always longer than that of the absorbed light. A solution of fluorescein in water will absorb the blue portion of white light and will fluoresce with light of a greenish hue. Thus a beam of white light traversing the solution becomes visible when observed from the side. Certain solids show a persistence of the reemitted light, so that it lasts several seconds or even minutes after the incident light is turned off. This is called *phosphorescence*.

Very striking fluorescent effects may be produced by illuminating various objects with ultraviolet light from a mercury arc. A special nickel oxide glass can be obtained which is almost entirely opaque to visible light but transmits freely the strong group of mercury lines near $\lambda 3650$. If only this light from the arc comes through the glass, many organic as well as inorganic substances are rendered visible almost exclusively by their fluorescent light. The teeth when illuminated by ultraviolet light will appear unnaturally bright, but artificial teeth look perfectly black.

22.7. Selective Reflection. Residual Rays. Substances are said to show selective reflection when certain wavelengths are reflected much more strongly than others. This usually occurs at those wavelengths for which the medium possesses very strong absorption. We are speaking now of dielectric substances, *i.e.*, those which are nonconductors of electricity. The case of metals is rather different and will be considered later in Chap. 28. That there is an intimate connection between selective reflection, absorption, and resonance radiation may be seen from an interesting observation made by R. W. Wood with mercury vapor. At a pressure of a small fraction of a millimeter, mercury vapor shows the phenomenon of resonance radiation when illuminated by $\lambda 2536$ from a mercury arc. As the pressure of the vapor is increased, the resonance radiation becomes more and more concentrated toward the surface of the vapor where the incident radiation enters, *i.e.*, on the inner wall of the enclosing vessel. Finally, at a sufficiently high pressure, the secondary radiation ceases to be visible except when viewed at an angle corresponding to the law of reflection. At this angle fully 25 per cent of the incident light is reflected in the ordinary way, the remainder having been absorbed and transformed into heat by atomic collisions. However, this high reflection, which is comparable to that of metals in this region, exists only for the particular wavelength $\lambda 2536$. Other wavelengths are

freely transmitted. In this experiment we evidently have a continuous transition from resonance radiation to selective reflection.

A few solids which have strong absorption bands in the visible region also show selective reflection. The dye fuchsine is an example. Such substances have a peculiar metallic sheen by reflected light and are strongly colored. Their color is due to the very high reflection of a certain band of wavelengths—so high that it is frequently termed “metallic” reflection. It is this type of reflection that is responsible for surface color (Sec. 22.1).

The most important application of selective reflection has been its use in locating absorption bands which lie far in the infrared. For example, quartz is found to reflect 80 to 90 per cent of radiation having a wavelength of about $8.5\ \mu$, or 85,000 Å. The method of *residual rays* for isolating a narrow band of wavelengths is based upon this fact. In Fig. 22E, S is a thermal source of radiation, giving a continuous spectrum. After reflection from the four quartz plates Q_1 to Q_4 , the radiation is

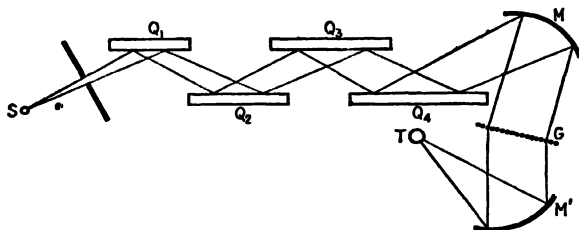


FIG. 22E. Experimental arrangement for observing residual rays by selective reflection.

analyzed by means of a wire grating G and thermopile T . It is found to consist almost entirely of the wavelength $8.5\ \mu$. Supposing this wavelength to be 90 per cent reflected at each quartz surface, and other wavelengths 4 per cent reflected, we have, after four reflections $(0.9)^4 = 0.66$ of the former remaining, but only $(0.04)^4 = 0.000026$ of the latter. The wavelengths of the residual rays of many substances have been measured in this way. Among the longest wavelengths measured are those from sodium chloride, potassium chloride, and rubidium chloride at $52\ \mu$, $63\ \mu$, and $74\ \mu$, respectively.

22.8. Theory of the Connection between Absorption and Reflection.

In Sec. 21.11 we mentioned briefly the mechanism postulated in the electromagnetic theory for the production of resonance radiation. It is assumed that light waves are incident upon matter which contains *bound charges* capable of vibrating with a natural frequency equal to that of the impressed wave. Thus a charge e is acted upon by the electric field E with a force eE , and if E varies with a frequency exactly

matching that with which the charged particle would normally vibrate, a large amplitude may be produced. As a result, the charged particle will reradiate an electromagnetic wave of the same frequency. In a gas at low pressure, where the atoms are relatively far apart, the frequency which can be absorbed will be sharply defined, and there will be no systematic relation between the phases of the light reemitted from different particles. The observed intensity from N particles will then be just N times that due to one particle (Sec. 12.4). This is the case with resonance radiation.

If, on the other hand, the particles are close together and interacting strongly with each other, as in a liquid or solid, the absorption will not be limited to a sharply defined frequency but will spread over a considerable range. The result is that the phases of the reemitted light from adjacent particles will agree. This will give rise to regular reflection, since the various secondary waves from the atoms in the surface will cooperate to produce a reflected wave front traveling off at an angle equal to the angle of incidence. In fact, this is just the conception used in applying Huygens' principle to prove the law of reflection. Hence selective reflection is also a phenomenon of resonance, and occurs strongly near those wavelengths corresponding to natural frequencies of the bound charges in the substance. The substance will not transmit light of these wavelengths; instead it reflects strongly. True absorption, or the conversion of the light energy into heat, may also occur to a greater or less extent because of the large amplitudes of the vibrating charges which are here involved. If absorption were entirely absent, the reflecting power would be 100 per cent at the wavelengths in question.

22.9. Scattering by Small Particles. The lateral scattering of a beam of light as it traverses a cloud of fine suspended matter was mentioned in Sec. 22.2. That this phenomenon is closely connected both with reflection and with diffraction may be seen by consideration of Fig. 22*F*. In (a) is shown a parallel beam consisting of plane waves advancing toward the right and striking a small plane reflecting surface. The successive wave fronts drawn are one wavelength apart, so that here the size of the reflector is somewhat greater than a wavelength. The light coming off from the surface of the reflector is produced by the vibration of the electric charges in the surface with a definite phase relation, and the spherical wavelets produced by these vibrations cooperate to produce short segments of plane wave fronts. These are not sharply bounded at their edges by the reflected rays from the edges of the mirror (dotted lines) but spread out somewhat, owing to diffraction. In fact, the distribution of the intensity of the reflected light with angle is just that derived in Sec. 15.2 for the light transmitted by a single slit. The

width of the reflector here takes the place of the slit width, so that we shall have greater spreading the smaller the width of the reflector relative to the wavelength.

In (b) of the figure, the reflector is much smaller than a wavelength, and here the spreading is so great that the reflected waves differ very little from uniform spherical waves. In this case the light taken from the primary beam is said to be scattered, rather than reflected, since the law of reflection has ceased to be applicable. Scattering is therefore a special case of diffraction. The wave scattered from an object much smaller than a wavelength of light will be spherical, regardless of whether or not the object has the plane form assumed in Fig. 22F(b). This follows from the fact that there can be no interference between the wavelets emitted by the several points on the surface of the scattering particle, inasmuch as the extreme points are separated by a distance much less than the wavelength.

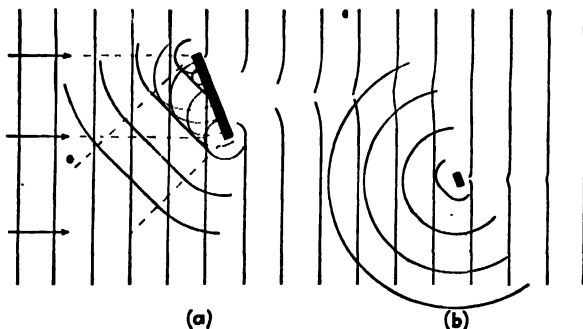


FIG. 22F. The reflection and diffraction of light by small objects comparable in size to the wavelength of light.

The first quantitative study of the laws of scattering by small particles was made in 1871 by Rayleigh,* and such scattering is frequently called *Rayleigh scattering*. The mathematical investigation of the problem gave a general law for the intensity of the scattered light, applicable to any particles of index of refraction different from that of the surrounding medium. The only restriction is that the linear dimensions of the particles be considerably smaller than the wavelength. As we might expect, the scattered intensity is found to be proportional to the incident intensity and to the square of the volume of the scattering particle. The most interesting result, however, is the dependence of scattering on wavelength. With a given size of the particles, long waves would be

* Several interesting papers laying the foundation of the theory will be found in "The Scientific Papers of Lord Rayleigh," Vols. 1 and 4, Cambridge University Press, New York, 1912.

expected to be less effectively scattered than short ones, because the particles present obstructions to the waves which are smaller *compared to the wavelength* for long waves than for short ones. In fact, it is found that the intensity is proportional to $1/\lambda^4$. Since red light, $\lambda 7200$, has a wavelength 1.8 times as great as violet light, $\lambda 4000$, the law predicts $(1.8)^4$ or 10 times greater scattering for the violet light from particles much smaller than the wavelength of either color. Figure 22G gives a quantitative plot of this relation.

If white light is scattered from sufficiently fine particles, such as those in tobacco smoke, the scattered light always has a bluish color. If the size of the particles is increased until they are no longer small compared to the wavelength, the light becomes white, as a result of ordinary diffuse reflection from the surface of the particles. The blue color seen with very small particles, and its dependence on the size of the particles,

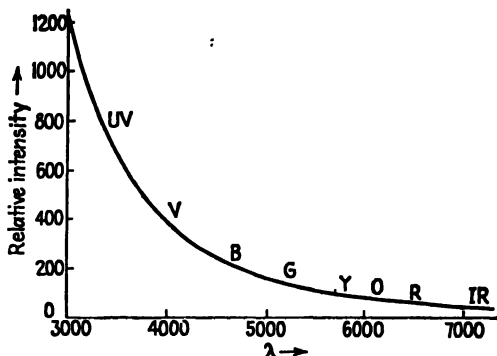


FIG. 22G. Intensity of scattering versus wavelength according to Rayleigh's law.

were first studied experimentally by Tyndall,* and his name is often associated with the phenomenon. Chalk dust from an eraser, falling across a beam of light from a carbon arc, will illustrate very effectively the white light scattered by large particles.

22.10. Molecular Scattering. Blue Color of the Sky. If a strong beam of sunlight is caused to traverse a pure liquid which has been carefully prepared to be as free as possible of all suspended particles of dust, etc., observation in a dark room will show that there is a small amount of bluish light scattered laterally from the beam. Although some of this light is still due to microscopic particles in suspension, which seem to be almost impossible to eliminate entirely, a certain amount

* John Tyndall (1820–1893). British “natural philosopher,” after 1867 superintendent of the Royal Institution and colleague of Faraday. Tyndall was outstanding for his ability to popularize and clarify physical discoveries.

appears to be attributable to the scattering by individual molecules of the liquid. At first sight it is surprising to find that the scattering from liquids is so feeble, in view of the large concentration of molecules present. It is, in fact, much weaker than the scattering from the same number of molecules of a gas. In the latter, the molecules are randomly distributed in space, and in any direction except the forward one the waves scattered by different molecules have perfectly random phases. Thus for N molecules the resultant intensity is just N times that scattered from any individual one (see Sec. 12.4). In a liquid, and even more so in a solid, the spacial distribution has a certain degree of regularity. Furthermore, the forces between molecules act to destroy the independence of phases (Sec. 22.8). The result is that the scattering from liquids and solids in directions other than forward is very weak indeed. The forward-scattered waves are strong and play an essential part in determining the velocity of light in the medium, as we shall see in the following chapter.

Lateral scattering from gases is also weak, but here the weakness is due to the relatively smaller number of scattering centers. When a great thickness of gas is available, however, as in our atmosphere, the scattered light is easily observed. It has been shown by Rayleigh that practically all the light that we see in a clear sky is due to scattering by the molecules of air. If it were not for our atmosphere, the sky would look perfectly black. Actually, molecular scattering causes a considerable amount of light to reach the observer in directions making an angle with that of the direct sunlight, and thus the sky appears bright. Its blue color is the result of the greater scattering of short waves. Rayleigh measured the relative amount of light of different wavelengths in sky light and found rather close agreement with the $1/\lambda^4$ law. The same phenomenon is responsible for the red color of the sun and surrounding sky at sunset. In this case, the scattering removes the blue rays from the direct beam more effectively than the red, and the very great thickness of the atmosphere traversed gives the transmitted light its intense red hue. An experiment demonstrating both the blue of the sky and the red of the sun at sunset is described in Sec. 24.15.

22.11. Raman* Effect. This is a scattering with change of wavelength somewhat similar to fluorescence. It differs from fluorescence, however, in two important respects. In the first place, the light which is incident on the scattering material must have a wavelength that does not correspond to one of the absorption lines or bands of the material. Otherwise we obtain fluorescence, as in the experiment of Sec. 22.5, where the green

* C. V. Raman (1888-). Professor at the University of Calcutta. He was awarded the Nobel prize in 1930 for his work on scattering and for the discovery of the effect that bears his name.

line of mercury is absorbed by the iodine vapor. In the second place, the intensity of the light scattered in the Raman effect is much less intense than most fluorescent light. For this reason the Raman effect is rather difficult to detect, and observations must usually be made by photography.

The apparatus illustrated in Fig. 22C is well adapted to observations of the Raman effect. For this purpose, a liquid or gas which is transparent to the incident light must be used in the tube *C*. It is convenient to fill tube *B* with a saturated solution of sodium nitrite, since this absorbs the ultraviolet lines of the mercury arc but transmits the blue-violet line $\lambda 4358$ with great intensity. Figure 22D(e) shows the Raman spectrum of CCl_4 . It will be seen that the same pattern of Raman lines is excited by each of the strong mercury lines. Figure 22D(d) illustrates the Raman spectrum of gaseous hydrogen, showing two sets of lines on the side toward the red of the exciting line, which in this case was $\lambda 2536$. Occasionally still fainter lines are seen on the violet side, two of which are visible in (d) and three in (e). This is also sometimes observed in the case of fluorescence. Since the modified light in these lines has a shorter wavelength than the incident light, they represent a violation of Stokes' law (Sec. 22.6) and are called *anti-Stokes* lines.

22.12. Theory of Scattering. When an electromagnetic wave passes over a small elastically bound charged particle, the particle will be set into motion by the electric force *E*. In Sec. 22.8 we considered the case where the frequency of the wave was equal to the natural frequency of free vibration of the particle. We then obtain resonance and fluorescence under certain conditions, and selective reflection under others. In both cases there may exist a considerable amount of absorption. Scattering, on the other hand, takes place for frequencies not corresponding to the natural frequencies of the particles. The resulting motion of the particles is then one of *forced vibration*. If the particle is bound by a force obeying Hooke's law, this vibration will have the same frequency and direction as that of the electric force in the wave. Its amplitude, however, will be very much smaller than that which would be produced by resonance. Hence the amplitude of the scattered wave will be much less, and this accounts for the relative faintness of molecular scattering. The phase of the forced vibration will differ from that of the incident wave, and this fact is responsible for difference of the velocity of light in the medium from that in free space. Thus scattering forms the basis of dispersion, which is to be discussed in the following chapter.

The electromagnetic theory is also capable of giving a qualitative picture of the changes of wavelength which occur in the Raman effect and in fluorescence. If the charged oscillator is bound by a force which does not obey Hooke's law, but some more complicated law, it will be capable

of reradiating not only the impressed frequency, but also various combinations of this frequency with the fundamental and overtone frequencies of the oscillator. For the complete explanation of these phenomena, however, the electromagnetic theory alone is not adequate. It cannot explain the actual magnitudes of the changes in frequency nor the fact that these are predominantly toward lower frequencies. For this, the quantum theory is required.

To investigate the distribution of intensity of the scattered light in different directions, let us consider the radiation from the charged particle e in Fig. 22H. Although Rayleigh scattering will occur from an uncharged particle as well, e may then be considered as an electron in one of its atoms. A plane-polarized electromagnetic wave with \mathbf{E} and \mathbf{H} in the directions indicated falls upon the particle, which then executes a forced vibration, parallel to the \mathbf{E} vector and having the frequency of the wave. This vibration of e generates a spherical wave about e as a

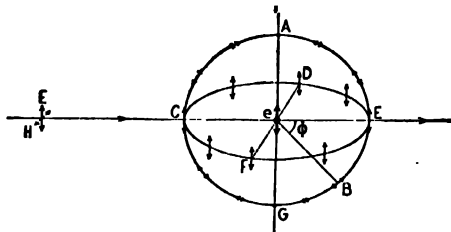


FIG. 22H. Showing the amplitude of the light wave scattered in all directions from a single electrically charged particle e .

center. But the amplitude of this wave is not constant over the wave front. Because an electromagnetic wave is necessarily a transverse wave, the vibration is in the wave front, perpendicular to the direction of travel. Now the vibration of e is incapable of producing any transverse disturbance along the direction eA or eG , and the intensity in these directions will be zero. Along eC the vibration has its maximum effect for transverse waves, and the same is true for all other points on the circle $CDEF$. In other directions such as eB the effective part of the vibration of e will be its projection perpendicular to this direction. If we call r_0 the maximum amplitude existing along $CDEF$, we have, for the amplitude of the wave in any direction making an angle ϕ with the plane containing $CDEF$,

$$r = r_0 \cos \phi \quad (22c)$$

Hence the intensity varies as $\cos^2 \phi$.

When the incident light is unpolarized, we must imagine the vectors \mathbf{E} oriented at random in all possible directions in a plane perpendicular to the direction of travel of the incident light. A little consideration will show that this leads to unpolarized light for the directions eC' and eE and completely plane-polarized light around the circle $AFGD$. These polarization effects predicted for the scattered light are completely confirmed by experiment and will be further discussed in Chap. 24.

22.13. Scattering and Refractive Index. The fact that the velocity of light in matter differs from that in vacuum is a consequence of scattering. The individual molecules scatter a certain part of the light falling on them, and the resulting scattered waves *interfere* with the primary wave, bringing about a change of phase which is equivalent to an alteration of the wave velocity. This process will be discussed in more detail in the chapter which follows, but here some simplified considerations may be used to show the connection between scattering and refractive index.

In Fig. 22I plane waves are shown striking an infinitely wide sheet of a transparent material, the thickness of which is small compared to the wavelength. Let the electric vector in this incident wave have unit amplitude, so that it may be represented at a particular time by $E = \sin 2\pi x/\lambda$. If the fraction of the wave that is scattered is small, the disturbance reaching some point P will be essentially the original wave, plus a small contribution due to the light scattered by all the atoms in the thin lamina. Now the energy scattered by a single atom is proportional to its *scattering cross section* σ , which is that part of the area presented by the atom to the oncoming waves which is effective in scattering these waves. The amplitude scattered from one atom is therefore proportional to $\sqrt{\sigma}$. If there are N atoms per cubic centimeter, the total scattered amplitude per square centimeter of surface becomes $E_s = \sqrt{\sigma} Nt$, where t is the thickness. These waves are all in phase as they leave the surface, since we have assumed $t \ll \lambda$. As the waves reach P , however, their phases will vary according to the different distances R from the different parts of the lamina. We may obtain the net effect by integrating over the surface, so that the total electric vector at P becomes

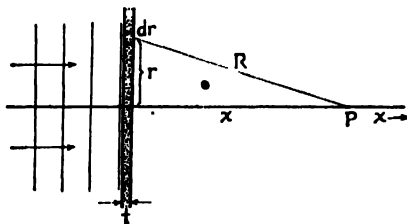


FIG. 22I. Geometry of scattering by a thin lamina.

$$E + E_s = \sin \frac{2\pi x}{\lambda} + \sqrt{\sigma} Nt \int_0^t \frac{2\pi r}{R} \sin \frac{2\pi R}{\lambda}$$

Now since $x^2 + r^2 = R^2$, and x is constant, we have $r dr = R dR$, and the integral may be written

$$\int_0^\infty \frac{2\pi}{R} \sin \frac{2\pi R}{\lambda} r dr = 2\pi \int_x^\infty \sin \frac{2\pi R}{\lambda} dR = 2\pi \frac{\lambda}{2\pi} \left[-\cos \frac{2\pi R}{\lambda} \right]_{R=x}^\infty$$

At $R = \infty$, it is legitimate to take the quantity in brackets as equal to zero, since one must assume at least a minute damping in order to avoid the existence of an infinitely long cosine wave, which is physically impossible (Sec. 11.3). Hence we have

$$E + E_s = \sin \frac{2\pi x}{\lambda} + \sqrt{\sigma} N t \lambda \cos \frac{2\pi x}{\lambda}$$

This is of the form $\sin A + B \cos A$, where B is assumed very small. Under this condition, one may write

$$\sin(A + B) = \sin A \cos B + \cos A \sin B \cong \sin A + B \cos A$$

Therefore

$$E + E_s = \sin \left(\frac{2\pi x}{\lambda} + \sqrt{\sigma} N t \lambda \right)$$

which shows that the phase of the wave at P has been altered by the amount $\sqrt{\sigma}/Nt\lambda$. But we know (Sec. 13.16) that the presence of a lamina of thickness t and refractive index n must retard the phase by $(2\pi/\lambda)(n - 1)t$. Hence

$$\sqrt{\sigma} N t \lambda = \frac{2\pi}{\lambda} (n - 1)t$$

and finally

$$n - 1 = \frac{1}{2\pi} N \lambda^2 \sqrt{\sigma}$$

This equation gives the relation between the index of refraction n and the scattering cross section σ . In this derivation no absorption has been considered, so that the equation is valid only for wavelengths well away from any absorption bands. In the next chapter we shall see how the refractive index behaves as the wavelength approaches that of an absorption band.

Problems

1. Are the residual rays from rock salt transmitted by sylvin?
2. What substance mentioned in this chapter could be used to take photographs using exclusively ultraviolet light?
3. Plot the intensity curve I against d for a medium having an absorption coefficient $\alpha = 0.32 \text{ cm}^{-1}$. Carry the curve as far as $I = I_0/10$.

4. If it has been determined that two-fifths of the apparent absorption in Prob. 3 is due to scattering, what would be the intensity of the beam, relative to I_0 , after traversing 5 cm of the medium, in case scattering were eliminated?

5. A solid substance possesses two absorption bands, each about 300 Å wide. One lies in the blue at $\lambda 4500$, and the other in the red at $\lambda 6000$. Taking the maximum value of the absorption coefficient for the first band to be 34 per cm, and of the second 230, draw curves of the distribution with wavelength for the light transmitted by thicknesses of 0.08 mm and 6 mm. Will the resultant color as seen by the eye be different in the two cases, and if so how? (NOTE: This effect is called *dichromatism*.)

6. A crystal reflects 60 per cent at the wavelength of its residual rays, and 5 per cent at adjacent wavelengths. How many reflections are necessary in order that the residual rays shall be 3000 times as strong as the light of adjacent wavelengths?

7. Compare the intensity of plane-polarized light scattered at 80° from the forward direction, in the plane containing the incident ray and the E vector, with the intensity scattered straight backward.

8. A small light source is just detectable by the eye at a distance R_0 when there is no scattering or absorption by the atmosphere. Derive an equation for the range R when the atmospheric transmission per mile is given by T . Assume the inverse square law to hold, and neglect the effect of the scattered light in the background on the visibility of the source.

9. Assuming the kinetic theory formula for the collision frequency, at what pressure of argon would the average time between collisions be equal to the average lifetime, 10^{-8} sec, of an excited argon atom? Assume a collision diameter of 3.36×10^{-8} cm for argon.

10. For sodium light, compute the approximate dimensions of the obstacles shown in Fig. 22F(a) and (b).

11. In photographing spectra of the Raman effect as excited by the resonance line of mercury, $\lambda 2536$, it is customary to place a few droplets of mercury in the body of the spectrograph in order to saturate the air in the light path with mercury vapor. Explain why this would be of advantage.

12. Make polar plots of the intensity of plane-polarized light scattered by bound electrons. Assuming the light to be incident in the $+x$ direction and the electric vector to lie along the y axis, make three plots: (a) for the x,y plane, (b) for the x,z plane, and (c) for the y,z plane.

CHAPTER 23

DISPERSION

The subject of dispersion concerns the velocity of light in material substances and its variation with wavelength. It is related to the angular dispersion of a prism mentioned in Sec. 2.10. This quantity is here to be represented as $d\theta/d\lambda$, the rate of change of the angle of deviation with wavelength. Now $d\theta/d\lambda$ may be considered as the product of two factors, $d\theta/dn$ and $dn/d\lambda$, n being the index of refraction of the prism. The first of these factors can be calculated from geometrical optics alone and depends on the refracting angle of the prism as well as on the angles of incidence on the first and second surfaces. The second factor is characteristic only of the medium of which the prism is made and is called the *dispersion* of the medium. It is the same, at a given wavelength, for any prism made of a given material. The dispersion, thus defined, is proportional to the rate of change of the reciprocal of the velocity with wavelength, since we have, by Eq. 11m, $n = c/v$, and

$$\frac{dn}{d\lambda} = \frac{d(c/v)}{d\lambda} = c \frac{d(1/v)}{d\lambda} \quad (23a)$$

Here, as throughout this chapter, the symbol λ represents the wavelength *in vacuo*. The fact that v depends on wavelength was first proved directly by Michelson's measurements of the velocity of red and blue light in carbon disulfide (Sec. 19.10). If, however, we accept the explanation of refraction given by the wave theory, the mere fact that at any refraction the different colors are refracted by different amounts necessitates a change of n , and therefore necessarily of v , with λ . In this chapter we shall first review some of the known facts about the variation of n with λ , and then inquire as to the physical reason for the corresponding changes in velocity.

23.1. Normal Dispersion. Suppose that the index of refraction of a glass prism has been measured for several different wavelengths of visible light. This can easily be done on a spectrometer by measuring the deviations by the prism of the various lines in a line spectrum. For prisms of some typical kinds of glass, one would obtain values like those shown in Table 23I. If any set of values of n are then plotted against

TABLE 23I. REFRACTIVE INDICES AND DISPERSIONS FOR SEVERAL COMMON TYPES OF OPTICAL GLASS

Wave-length, λ	Telescope crown		Borosilicate crown		Barium flint		Vitreous quartz	
	n	$\frac{dn}{d\lambda}$	n	$\frac{dn}{d\lambda}$	n	$\frac{dn}{d\lambda}$	n	$\frac{dn}{d\lambda}$
C 6563	1.52441	0.35×10^{-5}	1.50883	0.31×10^{-5}	1.58848	0.38×10^{-5}	1.45640	0.27×10^{-5}
6439	1.52490	0.36×10^{-5}	1.50917	0.32×10^{-5}	1.58896	0.39×10^{-5}	1.45674	0.28×10^{-5}
D 5890	1.52704	0.43×10^{-5}	1.51124	0.41×10^{-5}	1.59144	0.50×10^{-5}	1.45845	0.35×10^{-5}
5338	1.52989	0.58×10^{-5}	1.51386	0.55×10^{-5}	1.59463	0.68×10^{-5}	1.46067	0.45×10^{-5}
5086	1.53146	0.66×10^{-5}	1.51534	0.63×10^{-5}	1.59644	0.78×10^{-5}	1.46191	0.52×10^{-5}
F 4861	1.53303	0.78×10^{-5}	1.51690	0.72×10^{-5}	1.59825	0.89×10^{-5}	1.46318	0.60×10^{-5}
G 4340	1.53790	1.12×10^{-5}	1.52136	1.00×10^{-5}	1.60367	1.23×10^{-5}	1.46690	0.84×10^{-5}
H 3988	1.54245	1.39×10^{-5}	1.52546	1.26×10^{-5}	1.60870	1.72×10^{-5}	1.47030	1.12×10^{-5}

wavelength, a curve like one of those in Fig. 23A is obtained. The curves found for prisms of different optical materials will differ in detail but will all have the same general shape. These curves are representative

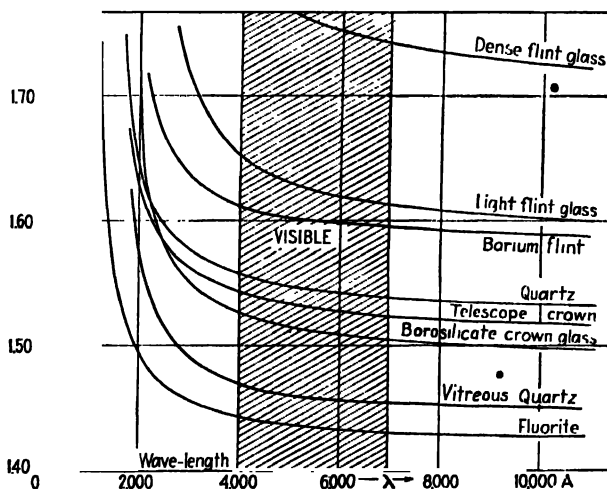


FIG. 23A. Dispersion curves for several different materials commonly used for lenses and prisms.

of *normal dispersion*, for which the following important facts are to be noted:

1. The index of refraction increases as the wavelength decreases.
2. The rate of increase becomes greater at shorter wavelengths.
3. For different substances the curve at a given wavelength is usually steeper the larger the index of refraction.

4. The curve for one substance cannot in general be obtained from that for another substance by a mere change in the scale of the ordinates.

The first of these facts agrees with the common observation that in refraction by a transparent substance the violet is more deviated than the red. The second fact can also be expressed by saying that the dispersion increases with decreasing wavelength. This follows because the dispersion, $dn/d\lambda$ is the slope of the curve (its negative sign is usually disregarded), which increases regularly toward smaller λ . An important consequence of this behavior of the dispersion is that in the spectrum formed by a prism the violet end of the spectrum is spread out on a much

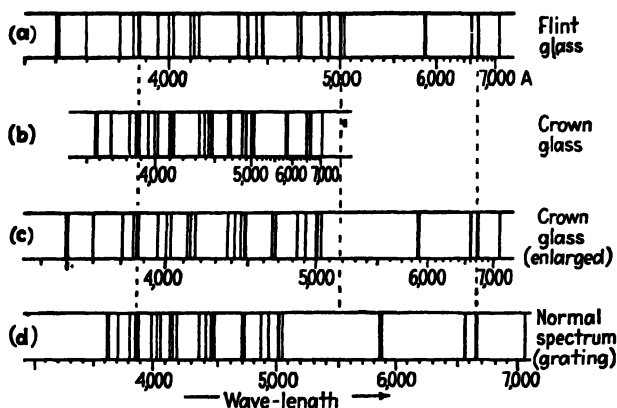


FIG. 23B. Comparison of the helium spectrum produced by flint- and crown-glass spectrographs with a normal spectrum.

larger scale than the red end. The spectrum is therefore far from being a normal spectrum (Sec. 17.6). This will be clear from Fig. 23B, in which the spectrum of helium is shown diagrammatically as given by flint- and crown-glass prisms and by a grating used under the proper conditions to give a normal spectrum. In the prism spectra the wave-length scale is compressed toward the red end, as can be seen by comparison with the uniform scale of the normal spectrum.

The third fact stated above requires that for a substance of higher index of refraction, the dispersion $dn/d\lambda$ shall also be greater. Thus, comparing (a) and (b) in Fig. 23B, the flint glass has the higher index of refraction, and gives a longer spectrum because of its greater dispersion. To compare the *relative* spacing of the lines in (b) with those in (a), the spectrum from crown glass has been enlarged, in (c), to have the same over-all length between the two lines λ_{3888} and λ_{6678} . When this is

done, it is seen that there is not complete agreement with the lines of (a). In fact, the spectra from prisms of different substances will never agree exactly in the relative spacing of their spectrum lines. This is a consequence of the fourth of the above facts, according to which the shape of the dispersion curve is different for every substance. The curve for flint glass in Fig. 23A has a greater slope at the violet end, relative to that in the red, than does the curve for crown glass. Consequently, the dispersion of different substances is said to be *irrational*, since there is no simple relation between the different curves.

All transparent substances which are not colored show normal dispersion in the visible region. The magnitude of the index of refraction may be quite different in various substances, but its change with wavelength always shows the characteristics described above. In general, the greater the density of the substance the higher its index of refraction and its dispersion. For example, flint glass has a density around 2.8, considerably higher than 2.4 for ordinary crown glass. Water has a smaller n and $dn/d\lambda$, while in a very light substance like air n is practically unity and $dn/d\lambda$ very nearly zero. For air $n = 1.000276$ for red light (Fraunhofer's C line), rising to only 1.000279 for blue light (F line). This rule relating density to index of refraction is only a qualitative one, and many exceptions are known. For instance, ether has a higher index than water (1.36 as compared with 1.33), yet it is less dense, as is shown by the fact that ether floats on the surface of water. Similarly, the correlation of high dispersion with high index is only rough, and there are exceptions to the third rule listed above. Diamond has a density of 3.52 and one of the highest known indices of refraction, varying from 2.4100 for the C line to 2.4354 for the F line. The difference in these values, which is a measure of the dispersion, is only 0.0254, whereas a dense flint glass may give as much as 0.0488 for the same quantity.

23.2. Cauchy's Equation. The first successful attempt to represent the curve of normal dispersion by an equation was made by Cauchy in 1836. His equation may be written

$$n = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4}$$

where A , B , and C are constants which are characteristic of any one substance. This equation represents the curves in the visible region, such as those shown in Fig. 23A, with considerable accuracy. To find the values of the three constants, it is necessary to know values of n for three different λ 's. Then three equations may be set up which, when solved as simultaneous equations, give A , B , and C . For some

purposes it is sufficiently accurate to include only the first two terms and the two constants can be found from values of n at only two λ 's. The two-constant Cauchy equation is, then,

$$n = A + \frac{B}{\lambda^2} \quad (23b)$$

from which the dispersion becomes, by differentiation of Eq. 23b,

$$\frac{dn}{d\lambda} = -\frac{2B}{\lambda^3} \quad (23c)$$

This shows that the dispersion varies approximately as the inverse cube of the wavelength. At 4000 Å it will be about eight times as large as

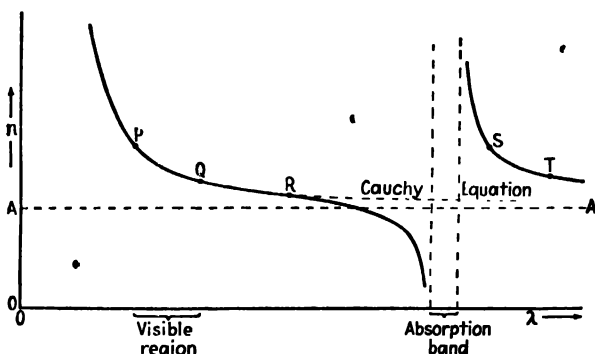


FIG. 23C. Anomalous dispersion of a transparent substance like quartz in the infrared. at 8000 Å. The minus sign means that the dispersion curve has a negative slope.

The theoretical reasoning on which Cauchy based his equation was later shown to be false, so that it is to be considered essentially as an empirical equation. Nevertheless it holds very satisfactorily for cases of normal dispersion and is a useful equation from a practical standpoint. We shall show later that it is a special case of a more complete equation, which has a sound theoretical foundation.

23.3. Anomalous Dispersion. If measurements of the index of refraction of a transparent substance like quartz are extended into the infrared region of the spectrum, the dispersion curve begins to show marked deviations from the Cauchy equation. The deviation is always of the type illustrated in Fig. 23C, where, starting at the point R, the index of refraction is seen to fall off more rapidly than required by a Cauchy equation that represents the values of n for visible light (between P and Q) quite accurately. This equation predicts a very gradual decrease of n for large values of λ (broken curve), the index approaching the

limiting value A as λ approaches infinity (Eq. 23b). In contrast to this, the measured value of n first decreases more and more rapidly as it approaches a region in the infrared where light ceases to be transmitted at all. This is an absorption band (Sec. 22.3), i.e., a region of selective absorption, the position of which is characteristic of the material. Within the absorption band, n cannot usually be measured because the substance will not transmit radiation of this wavelength. On the long-wavelength side of the absorption band the index is found to be very high, decreasing at first rapidly and then more slowly as we go farther beyond the absorption band. Over the range from S to T , the Cauchy equation will again represent the data, but with different constants. In particular, the constant A will be larger.

The existence of a large discontinuity in the dispersion curve as it crosses an absorption band gives rise to *anomalous dispersion*. The dispersion is anomalous because in this neighborhood the longer wavelengths have a higher value of n and are more refracted than certain shorter ones. The phenomenon was discovered with certain substances such as the dye fuchsine and iodine vapor whose absorption bands fall in the visible region. A prism formed of such a substance will deviate the red rays more than the violet, giving a spectrum which is very different from that formed by a substance having normal dispersion. When it was later discovered that transparent substances like glass and quartz possess regions of selective absorption in the infrared and ultra-violet, and therefore show anomalous dispersion in these regions, the term "anomalous" was seen to be inappropriate. No substance exists which does not have selective absorption at some wavelengths, and hence the phenomenon, far from being anomalous, is perfectly general. The so-called "normal" dispersion is found only when we observe those wavelengths which lie between two absorption bands, and fairly far removed from them. Nevertheless, the term "anomalous dispersion" has been retained, although it has little more than historical significance.

A very striking experiment showing the anomalous dispersion of sodium vapor in the neighborhood of the yellow D lines was devised by R. W. Wood in 1904. White light when passed through sodium vapor undergoes strong selective absorption at these lines, which form a close doublet of wavelengths 5890 and 5896 Å. At wavelengths far removed from these values, the index of refraction is only very slightly greater than unity, as we expect for a gas. With sodium vapor of appreciable density, the index of refraction in the neighborhood of the D lines passes through a region of anomalous dispersion (strictly speaking, two regions very

close together) of the type shown in Fig. 23C. As the D lines are approached from the side of shorter wavelengths, n begins to decrease rapidly, becoming much less than unity as we get very close to them. On the other side, it is at first very high, dropping off rapidly toward unity as λ increases further.

To show this in a direct way, Wood made use of the fact that we can produce the equivalent of a prism of sodium vapor by vaporizing the metal in a partially evacuated tube, if the tube is heated from the bottom. The arrangement is shown in Fig. 23D. A number of lumps of metallic sodium are placed along the bottom of a steel tube provided with water-cooled glass windows at the ends and an outlet for pumping. White light from a narrow horizontal slit S_1 is rendered parallel by the lens L_1 and after passing through the tube, forms a horizontal image S'_1 on the vertical slit S_2 of an ordinary prism spectroscope. When the sodium

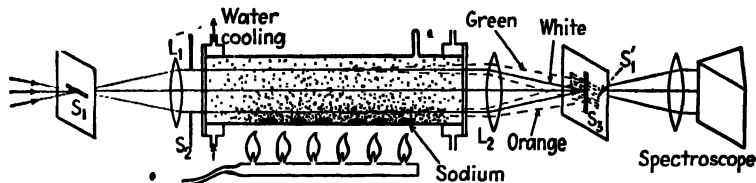


FIG. 23D. Experimental arrangement for observing the anomalous dispersion of sodium vapor.

tube is cold, S'_1 will be a sharp, white image, illuminating one point of the spectroscope slit, and this will be spread out into a narrow horizontal continuous spectrum in the focal plane of the spectroscope camera [Fig. 21H(a)]. If the tube is evacuated to about 2 cm pressure, and the sodium is heated by the row of gas burners, it will vaporize slowly, the vapor diffusing upward through the residual gas in the tube. A density gradient is set up, the vapor being densest at the bottom and rarest at the top of the tube. This is equivalent to a prism of vapor, the refracting edge of the prism being perpendicular to the plane of the figure, and its thickness increasing downward. This prism will form an anomalous spectrum on S_2 , in which the wavelengths shorter than the yellow (i.e., on the green side) are deviated upward, since their n is less than 1, and the longer ones (on the orange side) will be deviated downward. As a result, we might expect to observe in the spectroscope that the spectrum is deviated upward on the green side of the D lines, and downward on the red side. The directions are actually reversed because the spectroscope inverts the image of the slit. Three actual photographs of the resulting spectra with different densities of the vapor are shown in Fig. 23E. As a consequence of the inversion mentioned above, the

photographs form qualitatively a plot of n against λ , as in Fig. 23C. In the practical performance of this experiment, several refinements are desirable, of which an important one is the introduction of an auxiliary diaphragm S_2 to select that portion of the vapor where the density gradient is most uniform.*

23.4. Sellmeier's Equation. We have seen that the Cauchy equation is not capable of representing the dispersion curve in a region of anomalous dispersion. The first success in deriving a formula of more general applicability was obtained by postulating a mechanism by which the medium could affect the velocity of the light wave. It was assumed that the medium contains particles bound by elastic forces, so that they are capable of vibrating with a certain definite frequency ν_0 . This is the so-called *natural frequency*, *i.e.*, one with which the particles will vibrate in the absence of any periodic force, and is identical with the natural frequency mentioned in Sec. 22.8 in connection with absorption and selective reflection. Passage of the light waves through the medium is then assumed to exert a periodic force on the particles, which causes them to vibrate. If the frequency ν of the light wave does not agree with ν_0 , the vibrations will be forced vibrations of relatively small amplitude, and of frequency ν . As the frequency of the light approaches ν_0 , the response of the particles will be greater, and a very large amplitude will be built up by resonance when $\nu = \nu_0$ exactly. These vibrations will in turn react upon the light wave and alter its velocity: A mathematical investigation of this mechanism was made in 1871 by Sellmeier, who obtained the equation

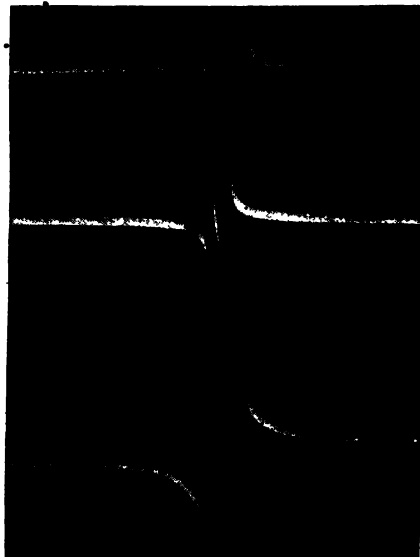


FIG. 23E. Anomalous dispersion of sodium vapor at three different gas densities. (After Cario.)

$$n^2 = 1 + \frac{A\lambda^2}{\lambda^2 - \lambda_0^2} \quad (23d)$$

This equation contains two constants, A and λ_0 , the latter being related to the natural frequency of the particles by the equation $\nu_0\lambda_0 = c$. Hence λ_0 is the wavelength in vacuum corresponding to ν_0 . To allow for the possibility of the existence of several different natural frequencies, the equation can be written with a series of terms,

$$n^2 = 1 + \frac{A_0\lambda^2}{\lambda^2 - \lambda_0^2} + \frac{A_1\lambda^2}{\lambda^2 - \lambda_1^2} + \cdots = 1 + \sum_i \frac{A_i\lambda^2}{\lambda^2 - \lambda_i^2} \quad (23e)$$

in which $\lambda_0, \lambda_1, \cdots$ correspond to the possible natural frequencies. The constants A_i are proportional to the number of oscillators capable of vibrating with these frequencies.

Figure 23*F* is a plot of n against λ according to Eq. 23*e*, assuming two natural frequencies. As λ approaches λ_0 or λ_1 , n goes to $-\infty$ or $+\infty$ on the short-wavelength or long-wavelength side, since the denominator

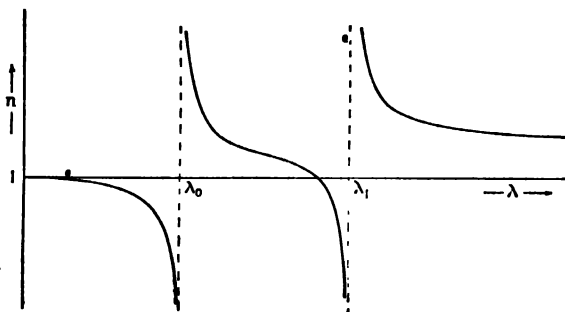


FIG. 23*F*. Theoretical dispersion curves given by Sellmeier's equation for a medium having two natural frequencies.

of one of the terms in Eq. 23*e* goes to zero. Other important characteristics of the curve to be noted are that n approaches unity as λ approaches zero, and that at $\lambda = \infty$, n^2 takes the value $1 + \sum_i A_i$.

Sellmeier's equation represents a great improvement over that of Cauchy and is in fact identical with that derived from the electromagnetic theory (Sec. 23.9) with certain simplifying assumptions. Not only does it take account of anomalous dispersion, but it also gives a more accurate representation of n in regions far from absorption bands than does a Cauchy equation with the same number of constants. That Cauchy's equation is an approximation to Sellmeier's can be seen by writing Eq. 23*d* in the form

$$n^2 = 1 + \frac{A}{\left(1 - \frac{\lambda_0^2}{\lambda^2}\right)}$$

On expanding by the binomial theorem, we find

$$n^2 = 1 + A \left(1 + \frac{\lambda_0^2}{\lambda^2} + \frac{\lambda_0^4}{\lambda^4} + \dots \right)$$

For that part of the dispersion curve where λ is considerably greater than λ_0 , the higher powers of λ_0/λ will be small and may be neglected. This gives

$$n^2 = 1 + A + A \frac{\lambda_0^2}{\lambda^2}$$

Writing M for $1 + A$, and N for $A\lambda_0^2$,

$$n = (M + N\lambda^{-2})^{\frac{1}{2}} \quad .$$

Expanding again,

$$n = M^{\frac{1}{2}} + \frac{N}{2M^{\frac{1}{2}}\lambda^2} + \frac{N^2}{8M^{\frac{3}{2}}\lambda^4} + \dots$$

and neglecting higher powers of $1/\lambda$,

$$n = P + \frac{Q}{\lambda^2} + \frac{R}{\lambda^4}$$

This is Cauchy's equation as given in Sec. 23.2. .

An instructive experiment to illustrate the origin of dispersion can be performed with a simple pendulum, to the bob of which is attached a light rubber band. If the end of the rubber band is held in the hand and moved to and fro, a periodic force is exerted on the pendulum similar to the action of the light wave on one of the oscillators in the medium. If the frequency motion of the hand is very high compared to the natural frequency of the pendulum, the bob will remain practically motionless. This corresponds to a wave of high frequency and short wavelength, the velocity of which is practically uninfluenced by the presence of the oscillators. In Fig. 23*F* it will be seen that n approaches unity as λ approaches zero, so the velocity becomes the same as in free space.

If, now, the hand is moved with a frequency only slightly greater than that of the pendulum, it will be found that the pendulum swings 180° out of phase with the motion of the hand. Under these conditions, the rubber band is considerably stretched when the displacements of the hand and bob are in opposite directions and so exerts its maximum force on the hand, tending to pull it back to the central position. This corresponds to an increased restoring force on the ether which propagates the wave, and hence to an *increase* in the velocity of the ether wave (Sec. 11.4). Thus in Fig. 23*F'*, n becomes considerably less than 1 at

a wavelength slightly less than λ_0 . Finally when the frequency of motion of the hand is made less than the natural frequency, the pendulum will follow the hand, practically in phase with it. In this case the rubber band exerts only small forces on the hand, since the displacements of the pendulum are in the same direction. The forces are less than if the pendulum were at rest, so this corresponds to a decreased restoring force on the ether. The velocity of the wave is therefore *decreased*, and n is greater than one, on the long-wavelength side of λ_0 .

The large discontinuity in the dispersion curve at λ_0 is thus seen to be a consequence of the abrupt change of phase by 180° of the oscillator relative to the impressed vibration as the latter passes through a natural frequency. This effect may be demonstrated very directly by hanging three pendulums side by side from a horizontal rod clamped at one end. The center pendulum is a heavy one and corresponds to the ether wave while the other two are very light, one being slightly longer and the other slightly shorter than the heavy pendulum. When the center pendulum is set swinging, the two light ones will swing in opposite phases, the shorter one nearly agreeing in phase with the impressed vibration.

23.5. Effect of Absorption on Dispersion. Although Sellmeier's equation represents the dispersion curve very successfully in regions not too close to absorption bands, it fails completely at those wavelengths where the medium has appreciable absorption. This may be seen directly from the fact that the curve of Fig. 23*F* goes to infinity on either side of each λ_i . Not only is this physically impossible, but the form of the curve near λ_i does not agree with experiment. It has been possible to measure the dispersion curve right through an absorption band, although this is a difficult matter because practically all the light is absorbed. By using prisms of very small refracting angle, or thin films of the material with a Michelson interferometer (Sec. 13.16), the indices of refraction of a few dyes, such as cyanine, which have an absorption band in the visible, have been carefully measured. The resulting curve resembles that shown by the heavy solid line in Fig. 23*G*. The true form of the curve in the neighborhood of λ_i is seen to be very different from that required by Sellmeier's equation (Fig. 23*F*).

This discrepancy was first shown by Helmholtz* to be due to the fact that Sellmeier's equation takes no account of the absorption of energy

* H. L. F. von Helmholtz (1821-1894). German physicist who contributed in almost every field of science. His work in physiological optics alone, or in sound, would have made him famous. He is regarded as one of the discoverers of the law of conservation of energy.

of the wave. In the above discussion, and in the suggested mechanical analogy, it was assumed that the oscillator does not experience any frictional resistance to its vibration. Such a resistance is necessary if energy is to be taken continuously from the wave by the oscillator. Helmholtz assumed a frictional force directly proportional to the velocity of the oscillator, and he therefore derived an equation for the index of refraction which takes account of absorption. As a measure of the

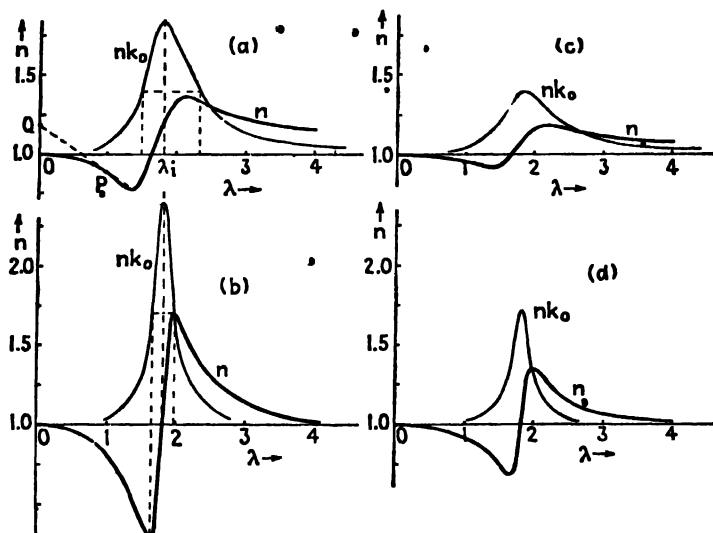


FIG. 23G. Ideal dispersion curves for an oscillator with different amounts of friction and absorption; (a) strong absorption, strong friction, (b) strong absorption, weak friction, (c) weak absorption, strong friction, and (d) weak absorption, weak friction.

strength of the absorption, we could use the absorption coefficient α defined in Eq. 11v, but the equations are simpler when expressed in terms of the extinction coefficient κ_0 , which is related to α as follows:

$$\kappa_0 = \frac{\alpha\lambda}{4\pi} \quad (23f)$$

Here λ is the wavelength measured in vacuum. The physical significance of κ_0 is best expressed by the fact that the intensity falls to $1/e^{4\pi\kappa_0}$ of its initial value in going the distance λ through the medium. The dispersion equation resulting from this purely mechanical theory of Helmholtz may be written

$$n^2 - \kappa_0^2 = 1 + \sum_i \frac{A_i \lambda^2}{(\lambda^2 - \lambda_i^2) + g_i \lambda^2 / (\lambda^2 - \lambda_i^2)} \quad (23g)$$

The constant g_i is a measure of the strength of the frictional force. This equation now holds for all wavelengths, including those within an absorption band. In regions far from absorption bands, κ_0 and g_i are both essentially zero, and the equation reduces to Sellmeier's equation 23*e*. Figure 23*G(a)* is a plot of n and of $n\kappa_0$, the latter of which by Eq. 23*f* is a measure of the absorption coefficient α , for a case of large friction ($g = 1.96 \times 10^{-3}$). It shows quantitatively the course of dispersion and absorption curves through a region of absorption with a maximum at $\lambda_i = 0.1732$ microns. It will be seen that n no longer goes to infinity, as in Fig. 23*F*, but remains finite at $\lambda = \lambda_i$. The other curves of Fig. 23*G* are drawn to show the effects of changing both the strength of the absorption and the amount of frictional loss. It should be noted in (b) and (d) that the maxima and minima of the refractive index curves come exactly at the points where the absorption is half its maximum value.

The pendulum experiments described above may be modified to include the effect of frictional damping and to give some insight into the physical reason for the resulting change in the form of the dispersion curve. Thus if the smaller pendulum, which represents the oscillator, has a wire attached to it which dips in water or oil, we have the desired condition. Two important changes in the response of the pendulum to the impressed vibrations will now be apparent. In the first place, the amplitude will not become nearly so large when the impressed frequency is exactly equal to the natural frequency of the pendulum. With no friction, the amplitude produced by resonance is theoretically infinite (in the final equilibrium state), and the corresponding value of n goes to infinity also. The effect of friction, however, limits this maximum amplitude, and this accounts for the fact that only moderate variations of n are actually observed. In the second place, the change of relative phase when the impressed vibrations pass through the natural frequency is no longer abrupt, but takes place more or less gradually. This accounts for the fact that there is no longer a sharp discontinuity in the dispersion curve, but that it is rounded off into a continuous curve. The phase change becomes more and more gradual as the friction is made greater, for instance by dipping the wire farther into the water, or by using a more viscous liquid.

23.6. Wave and Group Velocity in the Medium. In the curves of Figs. 23*F* and 23*G*, the abscissas are wavelengths in vacuum $\lambda = c/\nu$ and the ordinates are the ordinary indices of refraction $n = c/v$, where v is

the wave velocity in the medium. For those parts of the curve where $n < 1$, the wave velocity is greater than c , the velocity of light in vacuum. This is at first sight a contradiction to one of the fundamental results of the theory of relativity, according to which c is the highest attainable velocity. There is actually no contradiction here, however, since relativity applies to the velocity with which *energy* is transmitted, and this, as we shall show, is always less than c . The rate at which energy is carried by waves is determined by the group velocity u , which is related to the wave velocity by Eq. 12o:

$$u = v - \lambda_m \frac{dv}{d\lambda_m}$$

Here λ_m is the wavelength in the medium. It is possible to find the value of the ratio c/u from a plot of n against λ by a simple geometrical construction similar to that used in Sec. 12.8 to find u from a plot of v against λ . Thus, dividing Eq. 12o by c and taking the reciprocal, we have

$$\frac{c}{u} = \frac{c}{v - \lambda_m \frac{dv}{d\lambda_m}} = \frac{1}{\frac{1}{n} - \frac{\lambda}{n} \frac{d(1/n)}{d(\lambda/n)}}$$

since $v = c/n$ and $\lambda_m = \lambda/n$. This expression can be simplified to

$$\frac{c}{u} = n - \lambda \frac{dn}{d\lambda} \quad (23h)$$

By analogy with the treatment given in Sec. 12.8, it will be seen that if at a point P [Fig. 23G(a)] a tangent be drawn to the dispersion curve, it will intersect the axis of n at a point Q whose ordinate is c/u . That is, while the ordinate of P is n , or c/v , for that wavelength, the ordinate of Q is the corresponding value of c/u for the same wavelength.

This geometrical construction shows, then, that for any point on the curve where it is descending toward the right, the corresponding c/u is greater than unity, even though n itself may be less than unity. Therefore the group velocity is less than c , and there is no violation of the principle of relativity. An exception to this statement appears to occur in the region within the absorption band, where the curve slopes up steeply to the right. However, in this region we have strong absorption, so that the amplitude of the wave drops practically to zero in a fraction of a wavelength. In this event, the wave velocity and group velocity no longer have any meaning, but other considerations show that in this case also the relativity requirement is fulfilled.

23.7. The Complete Dispersion Curve of a Substance. Although the curve of the refractive index against wavelength is different for every different substance, the curves for all optical media, *i.e.*, substances more or less transparent in the visible region, possess certain general features in common. To illustrate these, let us consider the schematic curve of Fig. 23H, which represents the variation of n from $\lambda = 0$ to several kilometers for an ideal substance. Starting at $\lambda = 0$, the index of refraction is unity, as stated in Sec. 23.4. For the very short waves (γ rays and hard X rays), the index is slightly less than 1. Siegbahn* proved this experimentally by refracting X rays through a prism. It was found that the beam was deflected very slightly *away from* the base of the prism, as would be the case if the waves travel faster in the prism than in air. It has also been demonstrated that X rays can be totally reflected from a solid substance by using grazing incidence so that the X rays strike the surface at an angle greater than the critical angle.

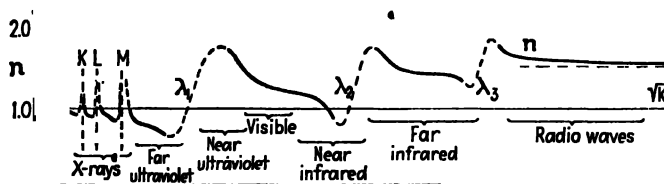


FIG. 23H. Schematic diagram of a complete dispersion curve for a substance transparent to the visible spectrum.

This property of X rays has been used by A. H. Compton† and others to measure the wavelengths of X rays by diffracting them from an ordinary ruled grating used at grazing incidence.

The first absorption is encountered in the X-ray region at a wavelength depending upon the atomic weight of the heaviest element in the material. For silicon it reaches its maximum at 6.731 Å, and for uranium at 0.1075 Å. This absorption rises rapidly to a maximum, and then falls off sharply at the so-called "K-absorption limit" of the element. It gives rise to a relatively narrow region of strong anomalous dispersion, marked K in Fig. 23H. Beyond this will lie other absorption discontinuities of this element, called L, M, . . . limits, as well as the K, L, M, . . . limits of other elements present. Therefore for any actual

optical medium there will be many of these sharp discontinuities. For simplicity only three are indicated in the figure.

From the X-ray region the curve descends more rapidly toward longer wavelengths, eventually reaching a broad region of strong absorption and anomalous dispersion in the ultraviolet (Sec. 22.3). For most optical media this completely covers the region between the soft X rays and the near ultraviolet. The descending course of the curve in the visible region, characteristic of normal dispersion, is seen to be connected with the existence of this ultraviolet absorption. In general the curve will have a steeper slope in the visible region, so that the dispersion $dn/d\lambda$ is greater, the nearer this absorption band lies to the visible. Thus fluorite has a very small dispersion for visible light, quartz somewhat greater, and glass still greater (cf. Fig. 23A and Table 22I). Dense flint glass, which gives the highest dispersion, frequently has a yellowish color, owing to the fact that the absorption band encroaches slightly into the violet end of the visible.

Somewhere in the near infrared, the curve begins to descend more steeply, and runs into another absorption band. The center of this band is at 8.5μ for quartz, but the absorption begins to become strong at 4 or 5μ . Beyond this first absorption band there usually exist one or more others. In passing each of these bands, the index of refraction increases. Thus the index will be higher for certain infrared wavelengths than for any part of the visible. For example, Rubens measured values of n for quartz varying from 2.40 to 2.14 in the region $\lambda = 51$ to 63μ . An interesting method of isolating radiation of very long wavelengths, called the method of "focal isolation" is based on this fact. Owing to the high value of n , a convex lens will have a much smaller focal length for these long waves than for the shorter waves, and the latter can be screened off with suitable diaphragms. In this way the longest infrared rays ever measured were isolated by Nichols and Tear (Sec. 11.5).

At wavelengths beyond all the infrared bands, the index decreases slowly and uniformly through the region of radio waves, approaching a certain limiting value for infinitely long waves. This value will be shown in the following section to be the square root of k , the ordinary dielectric constant of the medium.

23.8. The Electromagnetic Equations for Transparent Media. In Chap. 20 we stated Maxwell's equations as they apply to empty space, and we showed how they predict electromagnetic waves of velocity c . It is now of interest to investigate the characteristics and velocity of such waves in a material substance. For the present we shall consider only nonconducting media, and the more difficult case of conductors will be taken up later in Chap. 28. When a steady electric field acts upon a

nonconducting dielectric, there is a small displacement of the bound charges in the atoms, and we say they become *polarized*. The charges do not move continuously along, as in a conductor, but are merely displaced through minute distances and come to rest again in a fashion analogous to the stretching of a spring. As a measure of this *electric displacement* we use the vector quantity \mathbf{D} ,* and since in an isotropic medium it is proportional to the impressed electric field \mathbf{E} , we may write

$$\mathbf{D} = k\mathbf{E} \quad (23i)$$

Here k is the dielectric constant. To apply Maxwell's equations to such a medium it now becomes necessary to replace \mathbf{E} by \mathbf{D} wherever it occurs in the equations for empty space (Eqs. 20a to 20d). Hence Maxwell's equations for a nonconducting isotropic medium are written:

$$\left. \begin{aligned} \frac{k}{c} \frac{\partial E_x}{\partial t} &= \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} \\ \frac{k}{c} \frac{\partial E_y}{\partial t} &= \frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} \\ \frac{k}{c} \frac{\partial E_z}{\partial t} &= \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} \end{aligned} \right\} \quad (23j)$$

$$\left. \begin{aligned} -\frac{1}{c} \frac{\partial H_x}{\partial t} &= \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} \\ -\frac{1}{c} \frac{\partial H_y}{\partial t} &= \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} \\ -\frac{1}{c} \frac{\partial H_z}{\partial t} &= \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} \end{aligned} \right\} \quad (23k)$$

and

$$k \left(\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} \right) = 0 \quad (23l)$$

$$\frac{\partial H_x}{\partial x} + \frac{\partial H_y}{\partial y} + \frac{\partial H_z}{\partial z} = 0 \quad (23m)$$

These same equations are sometimes given in terms of the displacement current \mathbf{j} , the x component of which is written

$$j_x = \frac{k}{4\pi c} \frac{\partial E_x}{\partial t}$$

The left-hand member of each equation in Eq. 23j is therefore written as $4\pi j_x$, $4\pi j_y$, and $4\pi j_z$. They may also be written in terms of the displacement \mathbf{D} as is done in Eqs. 25e.

If we now derive the equations for plane waves as was done in Sec. 20.4, starting with Eqs. 23j and 23k, we find

$$\frac{\partial^2 H_x}{\partial t^2} = \frac{c^2}{k} \frac{\partial^2 H_x}{\partial x^2}$$

* Strictly speaking, \mathbf{D} itself is not a direct measure of the displacement of the bound charges. The polarization of the medium is usually written \mathbf{P} , and \mathbf{D} depends on \mathbf{P} by the relation $\mathbf{D} = \mathbf{E} + 4\pi\mathbf{P}$.

and

$$\frac{\partial^2 E_y}{\partial t^2} = \frac{c^2}{k} \frac{\partial^2 E_y}{\partial x^2}$$

Comparing with Eq. 20m, we see that the velocity of the waves is now c/\sqrt{k} . The index of refraction becomes

$$n = \frac{c}{v} = \sqrt{k} \quad (23n)$$

The solution of Eqs. 23j to 23m for monochromatic plane waves, analogous to Eqs. 20o, is now to be written,

$$E_y = E_y^0 \cos \frac{2\pi}{\lambda} (x - vt)$$

$$H_z = \sqrt{k} E_y^0 \cos \frac{2\pi}{\lambda} (x - vt)$$

and we have, for the relation between the amplitudes of the magnetic and electric waves,

$$H_z^0 = \sqrt{k} E_y^0$$

Therefore in the usual case $k > 1$ the amplitude of the magnetic wave is greater than that of the electric wave in a ratio equal to the index of refraction (Eq. 23n).

As regards the energy and intensity of the waves, the energy per unit volume of the electric wave is now found to be $kE_y^2/8\pi$, and of the magnetic wave $H_z^2/8\pi$. From the above relations these will be seen to be equal, so we may write

$$\text{Energy per unit volume} = \frac{kE_y^2}{4\pi} = \frac{H_z^2}{4\pi} = \frac{\sqrt{k} E_y H_z}{4\pi}$$

When this is multiplied by the velocity v from Eq. 23n, we get for the instantaneous rate of flow of energy across unit surface the quantity $cE_y H_z/4\pi$. This is the result, for the present case, of a very general law called *Poynting's* theorem*, according to which the rate of flow of energy is determined by the vector product of E and H . This instantaneous rate of flow is not the intensity however, for the intensity is determined by the rate of flow of energy per second, which is an *average* rate. This turns out to be proportional to the square of E_y^0 , the amplitude of the electric wave, and therefore also to the square of H_z^0 .

Equation 23n gives very nearly correct values of n for gases, but when we attempt to apply it to denser media, large deviations are found.

* J. H. Poynting (1852-1914). Professor of physics at the University of Birmingham, England. He is also known for his accurate work on the measurement of the gravitational constant.

Thus the dielectric constant for water, measured by placing it between the plates of a condenser charged to a steady potential, is 81, indicating a value of 9 for the index of refraction. For sodium light, the measured index of water is 1.33. For various kinds of glass, k varies from 4 to 9, which would require n to vary from 2 to 3. This again is higher than the observed values for visible light.

We do not have to look far for the cause of this discrepancy. It lies in the fact that the electric field of a light wave is not a steady field, but a rapidly alternating one. For yellow light the frequency is 5×10^{14} per sec. If the dielectric constant of a substance is measured using an alternating potential on the plates in place of a steady one, the result is found to vary with the frequency. From this we see that the index of refraction must also vary with frequency, or wavelength. As the wavelength becomes very large and approaches infinity, the frequency approaches zero. The limiting case of a steady field thus corresponds to zero frequency, and we are led to expect the index of refraction to approach the square root of the dielectric constant for steady fields. That this is in fact the case is shown by the measurements of the index of refraction of water for electromagnetic waves quoted in Table 23II.

TABLE 23II. VARIATION OF n WITH λ FOR WATER

Wavelength	Frequency	n
5.89×10^{-5} cm	5.1×10^{14}	1.333
12.56	2.9	1.3210
258.	0.116	1.41
800.	0.0375	1.41
0.40 cm	$750. \times 10^8$	5.3
1.75	171.	7.82
8.1	37.	8.10
65.	4.6	8.88
∞	0.	$(9.03 = \sqrt{k})$

The value of \sqrt{k} measured for steady field is shown for comparison. Clearly the value of n approaches exactly the predicted value for infinitely long waves.

23.9. Theory of Dispersion. In order to explain the variation of n (and hence of \sqrt{k}) with λ by the electromagnetic theory, one must take account of the molecular structure of matter. When an electromagnetic wave is incident on an atom or molecule, the periodic electric force of the wave sets the bound charges into a vibratory motion having the frequency of the wave. The phase of this motion relative to that of the

impressed electric force will depend upon the impressed frequency, and will vary with the difference between the impressed frequency and the natural frequency of the bound charges in the way discussed in Secs. 23.4 and 23.5. As the wave traverses the empty space between molecules, it will, of course, have the velocity c , and we must now inquire how it is possible that the presence of the oscillating charges in the molecules produces an effective alteration in the rate at which the wave progresses through the medium.

The clue to the explanation of dispersion lies in the secondary waves which are generated by the induced oscillations of the bound charges. These secondary waves are identical with those which give rise to molecular scattering (Sec. 22.12), as in the explanation of the blue color of the sky. When a light beam traverses a transparent liquid or solid, the amount of light scattered laterally is extremely small, even though the concentration of scattering centers is much greater than that in the air which gives the sky light. This is due to the fact that the scattered wavelets traveling out laterally from the beam have their phases so arranged that there is practically complete destructive interference. But the secondary waves traveling *in the same direction* as the original beam do not thus cancel out, but combine to form sets of waves moving parallel to the original waves. Now the secondary waves must be added to the primary ones according to the principle of superposition, and the results will depend on the phase difference between the two sets. This interference will modify the phase of the primary waves, and thus is equivalent to a change in their wave velocity. That is, since the wave velocity is merely the rate at which a condition of equal phase is propagated, an alteration of the phase by interference changes the velocity. We have seen that the phase of the oscillators, and hence of the secondary waves, depends on the impressed frequency, so it becomes clear that the velocity in the medium varies with the frequency of the light. This is the physical interpretation of dispersion, expressed in briefest outline.

The foundations for the mathematical treatment of the above mechanism were laid by Rayleigh, who considered the case of mechanical waves, and the theory was later extended to cover the case of electromagnetic waves by Planck, Schuster, and others. We shall not attempt to give this development here. It leads to a dispersion formula similar to that of Helmholtz (Eq. 23g). In fact, there is a close analogy throughout between the electromagnetic and mechanical pictures of the phenomenon. The oscillations of the bound charges must be regarded as damped by a frictional force, just as were the particles in Helmholtz's theory. On the electromagnetic theory, the damping is due to the radiation by the oscillator.

To show the relative amplitudes and phases of the incident wave, oscillator, and secondary wave, we consider the schematic diagrams of Fig. 23I. The first curve in (a) shows the response of a damped oscillator of natural frequency ν_0 to an impressed vibration of frequency ν . The amplitude becomes a maximum when $\nu = \nu_0$ and drops to zero at $\nu = \infty$. The broken curve shows the amplitude radiated by the oscillator, i.e., of the *scattered wave*. As a consequence of Rayleigh's law that the shorter waves are scattered more effectively, this curve is higher on the side of higher ν but drops to zero at low frequencies. The third curve gives the amplitude of the *secondary waves* built up from the scattered wavelets. Curve (b), in conjunction with the left-hand scale of ordinates, gives the phase difference between the oscillator and the impressed wave. As pointed out in Sec. 23.5, this changes from 0 to 180° in passing through

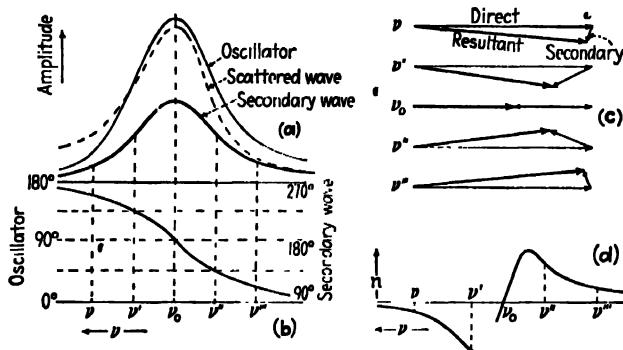


FIG. 23I. Illustrating the interpretation of dispersion as the result of interference of the secondary wave with the direct wave.

the natural frequency, but not abruptly because of the damping. At ν_0 it is 90° behind that of the impressed wave. Theory shows, furthermore, that the phase of the scattered waves, and therefore of the secondary waves as well, lags 90° behind that of the oscillators.* This is because electromagnetic radiation is proportional to the rate of change of current, or to the acceleration of a charge [see Sec. 20.8 and Fig. 20D(a)]. The current itself, or the velocity of the charge, has the phase that we attribute to the oscillator. Therefore, since in a simple periodic motion the acceleration is one-quarter period behind the velocity (Sec. 11.1), the phase of the radiated waves is retarded this much behind that of the oscillating source. Taking account of this additional retardation, it will be seen that the right-hand scale of ordinates in Fig. 23I(b) applies to the phase lag of the secondary waves behind the impressed waves.

We now proceed in (c) to compound vectorially the amplitudes of the direct and secondary waves. For the frequency ν , the amplitude of the secondary waves is small, [curve (a)] and lags in phase behind the direct waves by nearly 270° [curve (b)]. The vector diagram at the top in (c) shows that the resultant amplitude is nearly the same but that the phase is slightly *advanced*, corresponding to a rotation of the vector in a clockwise sense. An advance of phase means an *increase* in velocity, since it will be remembered that the phase increases as we move backward along a wave. Thus in the dispersion curve (d), the index of refraction at ν is slightly less than 1. The second vector diagram, for ν' , gives a greater advance of phase, and a considerably smaller resultant amplitude. The special case $\nu = \nu_0$ is interesting, since here there is no effect on the phase, and the velocity is the same as in free space: Note that $n = 1$ in curve (d). The smaller resultant amplitude at ν_0 , the center of the absorption band, is thus a consequence of interference. As in any case of interference, the intensity which is absent appears somewhere else, and here it is in the light scattered laterally. Beyond ν_0 , there is a retardation instead of an advance of phase, and the velocity of the wave is decreased. Thus it may be seen in a qualitative way how the curve (d), having the form for anomalous dispersion, can result from this mechanism.

23.10. Nature of the Vibrating Particles. In conclusion we may state very briefly the nature of the charged particles which are responsible for the discontinuities in the absorption and dispersion curves of a dielectric material (Fig. 23H). The discontinuities in the X-ray region (K , L , M , . . .) are known to be due to the innermost electrons in the atom. These are arranged in various "shells" of different energy, which are distinguished by the above capital letters. Because these electrons are well inside the atom, they are shielded from disturbing forces from other atoms, and this accounts for the relative sharpness of these absorption regions, even in solids.

The very broad absorption in the far ultraviolet is due to the outer electrons in the atoms and molecules of the material. These are not shielded, and consequently the region is broadened by the effects of neighboring atoms in solids or liquids. The near infrared absorption bands represent the various natural frequencies of the atoms as a whole, or even of molecules. Since these vibrators are much heavier than electrons, it is clear why they possess lower vibration frequencies. In the far infrared, other vibrations of lower frequency may be involved. Also the frequencies of rotation of molecules as a whole may operate, especially in gases.

Problems

1. Fit a three-constant Cauchy equation to the refractive indices listed for borosilicate crown glass in Table 23I. Calculate the constants so that the indices are exactly reproduced for $\lambda\lambda 6439$, 5338 , and 4340 , and report the results as a table of observed and calculated values, and the differences of these.

2. Investigate the inverse λ^2 dependence of the dispersion predicted by the two-constant Cauchy formula (see Eq. 23c). Test it in the cases of both barium flint glass and fused quartz, using the values of n listed in Table 23I for $\lambda\lambda 6563$ and $\lambda 3988$.

3. Using Cauchy's equation, evaluate the constants A and B for the six intervals given below. Average these values and plot a dispersion curve. Make a table of the correct and calculated n 's on the same graph paper.

$\lambda = 10000$	8000	6563	5349	4340	3404	$2763A$
$n = 1.5009$	1.5044	1.5074	1.5129	1.5209	1.5370	1.5603

(NOTE: Equation 23b should be used exclusively by solving two simultaneous equations for each interval.)

4. Using the dispersion formula derived from Cauchy's equation for the index of refraction, calculate the dispersion of a glass prism for (a) 4500 , (b) 5500 , and (c) 6500 Å. The constant $B = 1.6 \times 10^6 \text{ Å}^2$. Give units.

5. Using Cauchy's equation alone, calculate the dispersion of a glass prism at $\lambda_1 = 5000$. The constants $A = 1.420$ and $B = 1.8 \times 10^6 \text{ Å}^2$. Determine the value of n for two λ 's equidistant and on each side of λ_1 .

6. Assuming that Cauchy's equation is correct for a certain piece of glass, determine the dispersion at 5000 Å if the index of refraction $n = 1.53$ at 4000 Å and 1.48 at 5000 Å.

7. Repeat Prob. 4 using Sellmeier's equation, Eq. 23d.

8. Draw the dispersion curve for a substance which shows anomalous dispersion at 2500 , 3500 , and 6500 Å. The dielectric constant $k = 9$, the frictional term is the same for all, but the first resonance is twice as strong as the other two. Sketch the curve but do not attempt to calculate points.

9. Complete the derivation of Eq. 23k indicated in Sec. 23.6.

10. From Table 23I find the wave and group velocities of light of wavelength 5086 Å traveling in quartz.

11. The index of refraction of silver for X rays of wavelength 1.279 Å is 0.9999785 . Calculate the grazing angle (measured from the incident ray to the surface, in the plane of incidence) smaller than which total reflection will occur for X rays of this wavelength incident on a silver surface.

12. A crystal quartz lens has a focal length of 25 cm for sodium light. Calculate its focal length for a wavelength of 51μ . Draw to scale a diagram showing how the infrared light of this wavelength could be isolated from visible light by the method of focal isolation. See Sec. 23.7 and Table 25I for refractive indices.

13. Show the relative positions of the various visible spectral colors in the spectrum formed by a prism of a substance showing anomalous dispersion with the center of its band at $\lambda 5500$. Compare with the normal order of the colors in the spectrum from a glass prism.

14. Using the refractive indices from Table 23I, calculate the wave and group velocities for light of wavelength 5086 Å in barium flint glass.

15. In the case of relatively weak absorption, where κ_0 may be neglected as compared to n^2 in Eq. 23g, the half-maxima of the absorption curve occur at the same wave-

lengths as the maxima and minima of the dispersion curve. Determine graphically the width at half-maximum of an absorption band centered at 5000 Å, for which $g_i = 0.196 \times 10^6 \text{ Å}^2$.

16. Derive the wave equation for a dielectric, starting with Eqs. 23j and 23k, as suggested in Sec. 23.8.

17. A plane wave of amplitude E in vacuum enters a medium of index of refraction $n = 1.53$, striking the surface of the medium normally. Find the intensity of the light in the medium relative to that in vacuum.

CHAPTER 24

THE POLARIZATION OF LIGHT

From the properties of interference and diffraction we are led to conclude that light is a wave phenomenon, and we utilize these properties to measure the wavelength. These effects tell us nothing about the type of waves with which we are dealing—whether they are longitudinal or transverse, or whether the vibrations are linear, circular, or torsional. The electromagnetic theory, however, specifically requires that the vibrations be transverse, being therefore entirely confined to the plane of the wave front. The most general type of vibration is elliptical, of which linear and circular vibrations are extreme cases. Experiments which bring out these characteristics are those dealing with the so-called “polarization of light.” Although a longitudinal wave like a sound wave must necessarily be symmetrical about the direction of its propagation, transverse waves may show dissymmetries, and if any beam of light shows such a dissymmetry we say it is polarized.

The present chapter, by way of introduction to the subject of polarization, gives a brief account of the principal ways of producing plane-polarized light from ordinary unpolarized light. Most of the phenomena to be discussed here will be covered in more detail in later chapters. It will be helpful, however, to have a preliminary acquaintance with the experimental methods, and a mental picture of how the various polarizing devices act to separate ordinary light into its polarized components. The common methods used in producing and demonstrating the polarization of light may be grouped under the following heads: (1) reflection, (2) transmission through a pile of plates, (3) dichroism, (4) double refraction, and (5) scattering.

24.1. Polarization by Reflection. Perhaps the simplest method of polarizing light is the one discovered by Malus in 1808. If a beam of white light is incident at one certain angle on the polished surface of a plate of ordinary glass, it is found upon reflection to be plane-polarized. By “plane-polarized” is meant that all the light is vibrating parallel to a plane through the axis of the beam (Sec. 11.3). Although this light appears to the eye to be no different from the incident light, its polarization or asymmetry is easily shown by reflection from a second plate of glass as follows. A beam of unpolarized light, AB in Fig. 24A, is incident

at an angle of about 57° on the first glass surface at B . This light is again reflected at 57° by a second glass plate C placed parallel to the first as shown at the left. If now the upper plate is rotated about BC as an axis, the intensity of the reflected beam is found to decrease, reaching zero for a rotation of 90° . Rotation about BC keeps the angle of incidence constant. The experiment is best performed with the back surfaces of the glass painted black. The first reflected beam, BC' then appears to be cut off and to vanish at C' . As the upper mirror is rotated further about BC the reflected beam CD reappears, increasing in intensity to reach a maximum at 180° . Continued rotation produces zero intensity again at 270° , and a maximum again at 360° , the starting point.

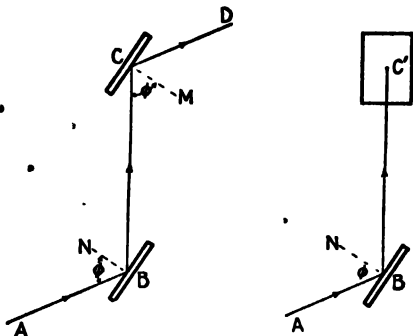


FIG. 24A. Polarization by reflection from glass surfaces.

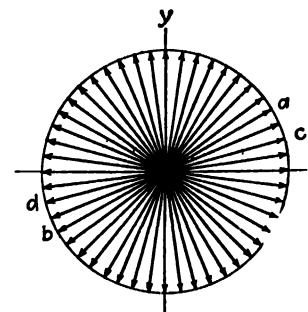


FIG. 24B. End view of vibrations of the electric vector, for an unpolarized beam of light.

If the angle of incidence on either the lower or upper mirror is not 57° , the twice-reflected beam will go through maxima and minima as before, but the minima will not have zero intensity. In other words there will always be a reflected beam from C . Calling the angle of incidence ϕ in general, the critical value ϕ which produces a zero minimum for the second reflection is called the *polarizing angle* and varies with the kind of glass used. Before explaining this experiment, which shows that the light reflected at the polarizing angle is plane-polarized, it should be pointed out that longitudinal waves like sound produce no such varying intensity when a similar experiment is performed with them.

24.2. Pictorial Representation of Light Vibrations. Let us assume that each light wave is a transverse wave whose vibrations are in straight lines at right angles to the direction of propagation. An ordinary beam of light, then, consisting of millions of such waves each with its own plane of vibration, would contain waves vibrating in all planes with equal probability. Looking at such a beam end-on (Fig. 24B), there would be an equal probability of finding a wave vibrating in the plane ab as there would be in any other plane cd .

The mode of vibration of any one light wave may be represented as shown in Fig. 24C(a), (b), (d), and (e). In (a) the light is traveling to the right and vibrating with the electric vector in the plane of the page.* The vibrations are represented by the short vertical lines. This same diagram may be taken to represent a beam of many light waves all vibrating parallel to the plane of the page. In (b) the vibrations are perpendicular to the page and the end-on view of the short lines repre-

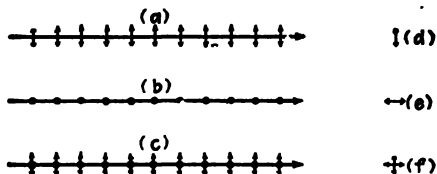


FIG. 24C. Pictorial representations of plane-polarized and ordinary light beams.

sents the vibrations are shown by dots. If there are two sets of waves vibrating at right angles to each other, they may be shown as in (c). End-on views of the same waves are represented in (d), (e), and (f).

It may now be shown that a beam of ordinary light vibrating in all planes may be thought of as consisting of two kinds of vibrations only, one set of waves vibrating in one plane as in (a) and another set vibrating at right angles as in (b). Consider any one of the light waves as for example ab shown in Fig. 24B. Let OA in Fig. 24D represent the amplitude of the electric displacement at an angle θ with two arbitrary axes x and y . Treated vectorially this is equivalent to two component vibrations, OA_x and OA_y , where

$$OA_x = OA \cos \theta \quad \text{and} \quad OA_y = OA \sin \theta \quad (24a)$$

FIG. 24D. Resolution of the amplitude of the electric vector of a light wave into components.

This process of resolution may be repeated for all vibrations in the ordinary light with the net result that the average of the amplitude components along the x axis will be just equal to the average along the y axis. Of course, these have no constant phase relation to each other. Thus Fig. 24C(c) and (f) may be taken to represent ordinary light. For a treatment of the average amplitude of many waves with random phases, see Sec. 12.4.

* The electromagnetic theory (Chap. 20) shows that there exist both an electric and a magnetic displacement to be associated with each light wave. Experiments to be described later (Sec. 28.11) show quite definitely that it is the electric vector that produces the observed optical effects with which we are familiar in polarized light.

24.3. Polarizing Angle and Brewster's Law. Consider unpolarized light to be incident at an angle ϕ on a dielectric like glass, as shown in Fig. 24E(a). There will always be a reflected ray OR and a refracted ray OT . An experiment like the one described in Sec. 24.1, and shown in Fig. 24A, shows that the reflected ray OR is partially plane-polarized and that only at a certain definite angle, about 57° for ordinary glass, is it plane-polarized. It was Brewster who first discovered that at this polarizing angle $\bar{\phi}$ the reflected and refracted rays are just 90° apart. This remarkable discovery enables one to correlate polarization with the refractive index

$$\frac{\sin \phi}{\sin \phi'} = n \quad (24b)$$

Since for the polarizing angle $\bar{\phi}$, $ROT' = 90^\circ$, the $\sin \bar{\phi}' = \cos \bar{\phi}$ giving

$$\frac{\sin \bar{\phi}}{\sin \bar{\phi}'} = \frac{\sin \bar{\phi}}{\cos \bar{\phi}} = n = \tan \bar{\phi} \quad (24c)$$

This is Brewster's law, which shows that the angle of maximum polarization depends on the refractive index and therefore varies with wavelength.

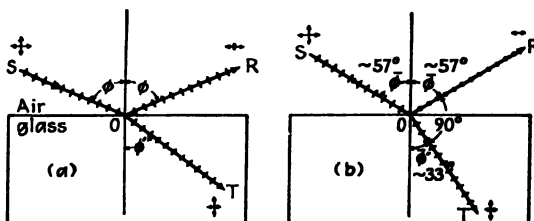


FIG. 24E. (a) Polarization by reflection and refraction. (b) Illustrating Brewster's law at the polarizing angle.

For ordinary glass the dispersion is such that the polarizing angle $\bar{\phi}$ over the whole visible spectrum does not vary appreciably. This is readily verified by referring to the dispersion curves in Fig. 23A and calculating $\bar{\phi}$ for several wavelengths. This is left as one of the problems at the end of the chapter.

The physical reason why light vibrating in the plane of incidence is not reflected at Brewster's angle lies in the transverse character of light vibrations. When the reflected beam travels at 90° with the refracted beam, the vibrations in the plane of incidence could generate only longitudinal waves traveling in the direction OR of Fig. 24E(b). Since such waves do not exist for light, there is zero reflection.

24.4. Polarization by a Pile of Plates. Upon examining the refracted light in Fig. 24E(a) for polarization, it is found to be partially polarized for all angles of incidence ϕ , with no angle for which the light is com-

pletely plane-polarized. The action of the reflecting surface may be described somewhat as follows: Let the ordinary incident light be thought of as being made up of two mutually perpendicular plane-polarized beams of light as shown in Sec. 24.2. Of those waves vibrating in the plane of incidence, i.e., in the plane of the page, part are reflected and part refracted for all angles with the single exception of the polarizing angle ϕ , for which all of the light is refracted. Of the waves vibrating perpendicular to the plane of incidence, some are always reflected without exception, and the rest refracted. Thus the refracted ray always contains some of both planes of polarization. For a single surface of glass with $n = 1.50$, it will be shown later [Sec. 28.1 and Fig. 28B(b)] that at

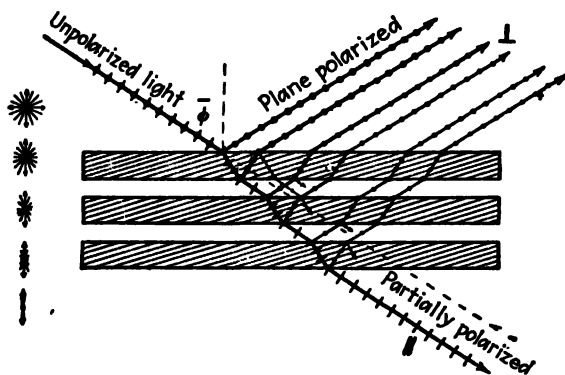


FIG. 24F. Polarization of light by a pile of glass plates.

the polarizing angle 100 per cent of the light vibrating parallel to the plane of incidence is transmitted, whereas for the perpendicular vibrations only 85 per cent is transmitted, the other 15 per cent being reflected. Obviously the degree of polarization of the transmitted beam is small for a single surface.

If a beam of ordinary light is incident at the polarizing angle on a pile of plates as shown in Fig. 24F, some of the vibrations perpendicular to the plane of incidence are reflected at each surface and all those parallel to it are refracted. The net result is that the reflected beams are all plane-polarized in the same plane,* and the refracted beam, having lost

* Special mention should be made of the fact that older books on optics refer to the *plane of polarization*, whereas newer books refer to the *plane of vibration*. Before polarization was well understood, light reflected from a dielectric at the polarizing angle was said to be polarized *in the plane of incidence*. Now it is known (Sec. 28.11) that optical phenomena are due to the action on matter of the electric vector and this is at right angles to the plane of incidence. It must be remembered, therefore, that the so-called plane of polarization is always perpendicular to the plane of vibration of the electric (light) vector.

more and more of its perpendicular vibrations, is partially plane-polarized. The larger the number of surfaces, the more nearly plane-polarized is this transmitted beam. This is illustrated by the vibration figures at the left in Fig. 24*F*. In a more detailed treatment of polarization by reflection and refraction (see Chap. 28), the polarizing angle for internal reflection is shown to correspond exactly to the angle of refraction ϕ' in Fig. 24*E(b)*. This means that light internally reflected at the angle ϕ' will also be plane-polarized.

Since a pile of plates forms a useful optical device for producing plane-polarized light, it is of importance to determine the percentage polarization of the light transmitted by any given number of plates. If I_p and I_s represent the intensity components of the light emerging from the plate which is vibrating parallel and perpendicular, respectively, to the plane of incidence, the *proportion of polarization* is defined by the relation

$$PP = \frac{I_p - I_s}{I_p + I_s} \quad (24d)$$

For a single surface, this amounts to $\frac{1}{185}$, or only 8.1 per cent. In deriving an equation for the magnitude of PP for a pile of plates, one must take into account the perpendicular components of light which by multiple reflection between the glass surfaces (see (Fig. 24*F*)) finally find their way through the last plate to be observed with the transmitted beam. The complete expression for the proportion of polarization, the derivation of which will not be given here, was first worked out in 1850 by Provostaye and Desains.* This equation is

$$PP = \frac{m}{m + \left(\frac{2n}{1 - n^2} \right)^2} \quad (24e)$$

where m is the number of plates (*i.e.*, $2m$ surfaces) and n the refractive index. For eight plates with $n = 1.50$ this gives a value of 0.582. If the corrections for multiple reflection are not made, one obtains a value of 0.857. This shows that multiple reflections, although they spoil the desired effect somewhat, must be taken into account (by the use of Eq. 24*e*) if a quantitatively correct result is desired.

A pile of plates to be used for polarizing light is usually mounted in a tube at such an angle that the light is incident on the plates at the

* Provostaye and Desains, *Ann. chim. et phys.*, **30**, 159, 1850. See also Geiger and Scheel, "Handbuch der Physik," Vol. 20, p. 217, Springer-Verlag, Berlin, 1928. It should be noted that most recent books on optics do not give the correct relations for the proportion of polarization produced by a pile of plates.

angle ϕ . Figure 24G shows two such piles, the polarizer (a) and the analyzer (b), with their planes of incidence parallel. The light emerging at N is nearly plane-polarized, and will be transmitted freely by the analyzer. Rotation of the latter by 90° about the line NM as an axis will cause the transmitted light to be nearly extinguished, since the vibrations are now perpendicular to the plane of incidence of the analyzer.

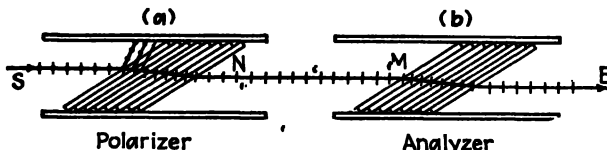


FIG. 24G. Glass plates mounted at the polarising angle ϕ .

and will be reflected to the side. A further rotation of 90° will restore the light, and in a complete revolution there will be two maxima and two minima. Any arrangement of polarizer and analyzer in tandem is called a *polariscope*, and as we shall see has numerous uses.

24.5. Law of Malus. This law tells us how the intensity transmitted by the analyzer varies with the angle that its plane of transmission makes with that of the polarizer. In the case of two piles of plates, the plane of transmission is the plane of incidence, and for the law of Malus to hold we must assume that the transmitted light is completely plane-polarized. A better illustration would be the double-reflection experiment of Sec. 24.1, or a combination of two polaroids or nicol prisms (see below), for which the polarization is complete. Then the law of Malus states that the transmitted intensity varies as the *square of the cosine* of the angle between the two planes of transmission.

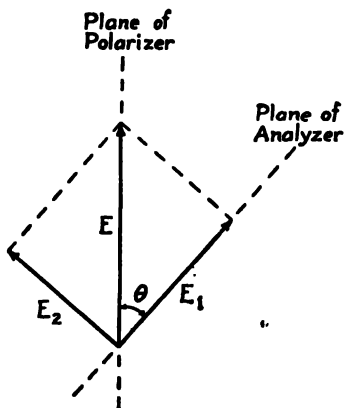


FIG. 24H. Resolution of the amplitude of plane-polarized light.

The proof of the law rests on the simple fact that any plane-polarized vibration—let us say the one produced by our polarizer—may be resolved into two components, one parallel to the transmission plane of the analyzer and the other at right angles to it. Only the first of these gets through. In Fig. 24H, let E represent the amplitude (electric vector) transmitted by the polarizer for which the plane of transmission intersects the plane of the figure in the vertical dashed line. When this light strikes the analyzer, set at the angle θ , one may resolve the incident amplitude

into components E_1 and E_2 , the latter of which is eliminated in the analyzer. In the pile of plates, it is reflected to one side. The amplitude of the light that passes through the analyzer is therefore

$$E_1 = E \cos \theta \quad (24f)$$

and its intensity

$$\begin{aligned} \bullet \quad I_1 &= E_1^2 = E^2 \cos^2 \theta \\ &= I_0 \cos^2 \theta \end{aligned} \quad (24g)$$

Here I_0 signifies the intensity* of the incident polarized light. This is, of course, one-half of the intensity of the unpolarized light striking the polarizer, provided one neglects losses of light by absorption in traversing it. There will also be losses in the analyzer. For polaroids or nicols there will be some light that is removed from the beam by reflection at the surfaces. Although these effects are neglected in deriving Eq. 24g, it will be noticed that they change only the constant in the equation and do not spoil the dependence of the *relative* intensity on $\cos^2 \theta$. Thus Malus' law is rigorously true and applies for example to the intensity of the twice-reflected light in the experiment of Sec. 24.1, even though its maximum value is only a small fraction of the original intensity. In such cases, the I_0 in Eq. 24g is merely the intensity when the analyzer is parallel to the polarizer.

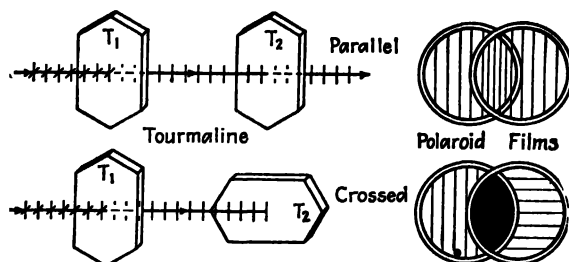


FIG. 24I. Diagram of tourmaline crystals and polaroid films illustrating polarization by selective absorption.

24.6. Polarization by Dichroic Crystals. These crystals have the property of selectively absorbing one of the two rectangular components of ordinary light. Dichroism is exhibited by a number of minerals and by some organic compounds. Perhaps the best known of the mineral crystals is *tourmaline*. When a pencil of ordinary light is sent through a thin slab of tourmaline like T_1 , shown in Fig. 24I, the transmitted light is found to be polarized. This can be verified by a second crystal T_2 . With T_1 and T_2 parallel to each other the light transmitted by the first crystal is also transmitted by the second. When the second crystal

is rotated through 90° , no light gets through. The observed effect is due to a selective absorption by tourmaline of all light rays vibrating in one particular plane (the O vibrations) but not those vibrating in a plane at right angles (the E vibrations). Thus in the figures shown, only the E vibrations parallel to the long edges of the crystals are transmitted so that no light will emerge from the crossed crystals. Since tourmaline crystals are somewhat colored, they are not used 'in optical instruments as polarizing or analyzing devices. They are useful, however, for those wavelengths of light for which they are transparent.

Attempts to produce polarizing crystals of large aperture were made by Herapath* in 1852. He was successful in producing good but small crystals of the organic compound iodosulfate of quinine (now known as herapathite or perapathite) which completely absorbs one component of polarization and transmits the other with little loss. One variety of the so-called *polaroid film* contains crystals of this substance. Polaroid was invented in 1932 by Land,† and has found uses in many different kinds of optical instruments. These films consist of thin sheets of nitrocellulose packed with ultramicroscopic polarizing crystals with their optic axes all parallel. In more recent developments the lining-up process is accomplished somewhat as follows: Polyvinyl alcohol films are stretched to line up the complex molecules and then are impregnated with iodine. From X-ray diffraction studies of these dichroic films, it can be seen that the iodine is present in polymeric form, *i.e.*, as independent long strings of iodine atoms all lying parallel to the fiber axis, with a periodicity in this direction of about 3.10 \AA . Films prepared in this way are called H-Polaroid. Land and Rogers found further that, when an oriented transparent film of polyvinyl alcohol is heated in the presence of an active dehydrating catalyst such as hydrogen chloride, the film darkens slightly and becomes strongly dichroic. Such a film becomes very stable and, having no dyestuffs, is not bleached by strong sunlight. This so-called K-Polaroid is very suitable for polarizing films for uses such as automobile headlights and visors.

24.7. Double Refraction. Up to the present time, the most useful method of producing and studying polarized light has been by double refraction in crystals of *calcite* and *quartz*. Both of these crystals are found in nature to be transparent to visible as well as to ultraviolet light. Calcite, which chemically is calcium carbonate, CaCO_3 , occurs in nature in a great variety of crystal forms (in the rhombohedral class of the hexagonal system), but it breaks readily into simple cleavage rhom-

bohedrons of the form shown at the left in Fig. 24J. Each face of the crystal is a parallelogram whose angles are $78^{\circ}5'$ and $101^{\circ}55'$. If struck a blow with a sharp instrument, each crystal may be made to cleave or break along cleavage planes into two or more smaller crystals which always have parallelogram faces with these same angles.

Quartz crystals, on the other hand, are found in their natural state to have many different forms, the most complicated of which is shown at the right in Fig. 24J. Unlike calcite, quartz crystals will not cleave

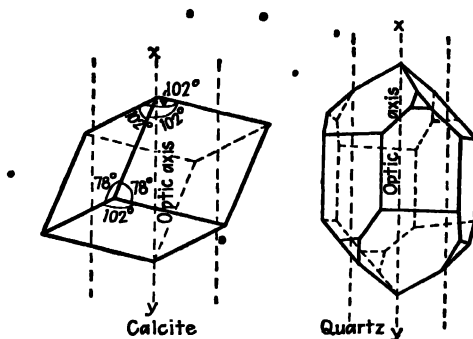


FIG. 24J. Calcite and quartz crystal forms. xy shows the direction of the optic axis.

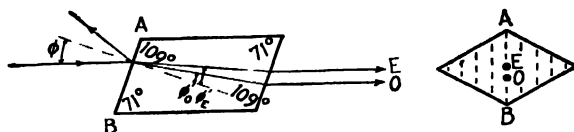


FIG. 24K. Illustrating double refraction of light in a calcite crystal.

along crystal planes but will break into irregular pieces when given a sharp blow. Their chemical constitution is silicon dioxide, SiO_2 . Further details concerning these crystals will be given in this as well as in the three following chapters.

When a beam of ordinary unpolarized light is incident on a calcite or quartz crystal, there will be, in addition to the reflected beam, two refracted beams in place of the usual one as in glass. This phenomenon, shown in Fig. 24K for calcite, is called *double refraction*. Upon measuring the angles of refraction ϕ' for different angles of incidence ϕ , one finds that Snell's law of refraction

$$\frac{\sin \phi}{\sin \phi'} = n \quad (24h)$$

holds for one ray but not for the other. The ray for which the law holds

is called the *ordinary* or *O ray* and the other is called the *extraordinary* or *E ray*.

Since the two opposite faces of a calcite crystal are always parallel, the two refracted rays emerge parallel to the incident beam and therefore parallel to each other. Inside the crystal the ordinary ray is always to be found in the plane of incidence. Only for special directions through the crystal is this true for the extraordinary ray.^f If the incident light is normal to the surface the extraordinary ray will be refracted at some angle that is not zero and will come out parallel to, but displaced from, the incident beam; the ordinary ray will pass straight through without deviation. A rotation of the crystal about the *O ray* will in this case cause the *E ray* to rotate around the fixed *O ray*.

24.8. Optic Axis. One important feature concerning the behavior of calcite and quartz is that there is one and only one direction through the crystal in which the *O* and *E rays* behave alike in all respects. There are other directions in which the double refraction, or separation of the rays, disappears, but only in this one direction do the velocities become alike as well (see Sec. 25.2 and Fig. 25*E*). This particular direction, called the *optic axis*, is shown by the dotted lines parallel to *xy* in Fig. 24*J*.

The direction of the optic axis in calcite is determined by drawing a line like *xy* through a *blunt corner* of the crystal so that it makes equal angles with all faces. A blunt corner is one where three obtuse face angles come together, and there are only two such angles that are opposite each other. It should be emphasized that the optic axis is *not* a line through the crystal but a *direction*. Through any given point in a crystal, however, there is one, and only one, line that can be drawn which makes equal angles with all faces and this is the optic axis for that point and all other points on that line.

In quartz the optic axis runs lengthwise of the crystal, its direction being parallel to the six side faces as shown. Just as in calcite, the optic axis is a direction through the crystal and not just a line. The importance of the optic axis in crystals will become apparent in what follows.

24.9. Principal Sections and Principal Planes. If a plane is passed through the optic axis and normal to a crystal surface, that plane is called a *principal section*. For every point there are therefore three principal sections, one for each pair of opposite crystal faces. A principal section always cuts the surfaces of a calcite crystal in a parallelogram with angles of 71° and 109° , as shown at the left in Fig. 24*K*. An end view of a principal section cuts the surface in a line parallel to *AB*, shown as a dotted line in the right-hand figure. All other planes through the crystal parallel to the plane represented by *AB* are also principal sections. These are represented by the other dotted lines.

The *principal plane of the ordinary ray* is defined as a plane in the crystal drawn through the optic axis and the ordinary ray, which can always be done, since the optic axis is merely a direction, and not a particular line, in the crystal. The *principal plane of the extraordinary ray* is defined as a plane in the crystal drawn through the optic axis and the extraordinary ray. The ordinary ray always lies in the plane of incidence. This is not generally true for the extraordinary ray. The principal planes of the two refracted rays do not coincide except in special cases. The special cases are those for which the plane of incidence is a principal section as shown in Fig. 24K. Under these conditions the plane of incidence, the principal section, and the principal planes of the *O* and *E* rays all coincide.

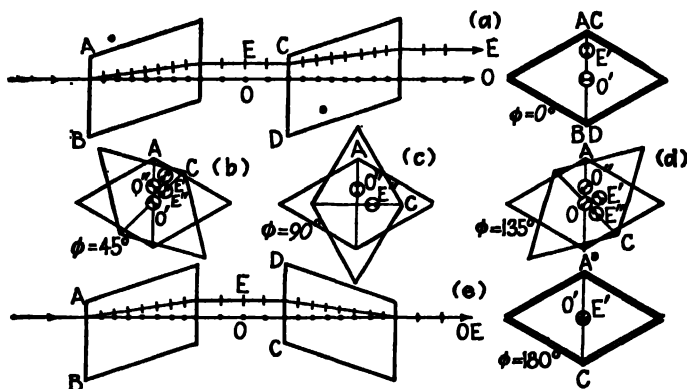


FIG. 24L. Double refraction and polarization in two calcite crystals with their principal sections making different angles.

24.10. Polarization by Double Refraction. The polarization of light by double refraction in calcite was discovered by Huygens in 1678. He sent a beam of light through two crystals as shown at the top in Fig. 24L. If the principal planes are parallel, the two rays *O'* and *E'* are separated by a distance equal to the sum of the two displacements found in each crystal if used separately. Upon rotating the second crystal each of the two rays *O* and *E* is refracted into two parts, making four as shown by an end-on view in (b). At 90° rotation the original *O'* and *E'* rays have faded and vanished and the new rays *O''* and *E''* have reached a maximum of intensity. Further rotation finds the original rays appearing and eventually, if the crystals are of equal thickness, these come together into one single beam in the center for the 180° position shown at the bottom, the rays *O''* and *E''* having now vanished.

Thus, merely by using two natural crystals of calcite, Huygens was able to demonstrate the polarization of light. The explanation of the

movement of the light rays is one simply of deviation by refraction and easily understood. The varying intensity of the spots, however, involves the polarization of the two light beams leaving the first crystal. In brief the explanation is somewhat as follows: Ordinary light upon entering the first calcite crystal is broken up into two plane-polarized rays, one, the

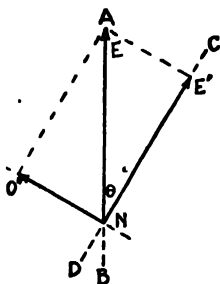


FIG. 24M. Resolution of polarized light into components by double refraction.

O ray, vibrating perpendicular to the principal plane, which is here the same as the principal section, and the other, the E ray, vibrating in the principal section. In other words, the crystal resolves the light into two components by causing one type of vibration to travel one path and the other vibration to travel another path.

Consider more in detail now what happens to one of the plane-polarized beams from the first crystal when it passes through the second crystal oriented at some arbitrary angle θ . Let E in Fig. 24M represent the amplitude of the E ray vibrating parallel to the principal section of the first crystal just as it strikes the face of the second crystal. This second crystal, like the first, transmits light vibrating in its principal section along one path and light vibrating at right angles along another path. The E ray is therefore split up into two components E' with an amplitude $E \cos \theta$ and O'' with an amplitude $E \sin \theta$. These emerge from the second crystal with relative intensities given by $E^2 \cos^2 \theta$ and $E^2 \sin^2 \theta$.

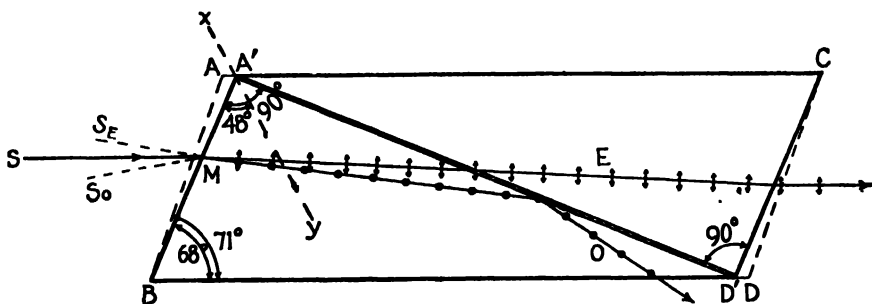


FIG. 24N. Detailed diagram of a nicol prism showing how it is made from a calcite crystal.

At $\theta = 90^\circ$ E' vanishes and O'' reaches a maximum intensity of A^2 . At all positions the sum of the two components, $E^2 \sin^2 \theta + E^2 \cos^2 \theta$, is just equal to E^2 the intensity of the incident beam.

The same treatment holds for the splitting up of the O beam from the first crystal into two plane-polarized beams O' and E'' .

24.11. Nicol Prism. This is an optical device made from a calcite crystal and is used in many optical instruments for producing and analyzing plane-polarized light. The nicol prism is made in such a way that it removes one of the two refracted rays by total reflection, as is illustrated in Fig. 24*N*. There are several different forms of nicol prism,* but we shall describe here one of the commonest ones. First a crystal about three times as long as it is wide is taken, and the ends cut down from 71° in the principal section to a more acute angle of 68° . The crystal is then cut apart along the plane $A'D'$ perpendicular to both the principal section and the end faces. The two cut surfaces are ground and polished optically flat and then cemented together with Canada balsam. Canada balsam is used because it is a clear transparent substance with an index of refraction about midway between the index of the O and E rays. For sodium yellow light, $\lambda 5893$,

Index of O ray.....	$n_o = 1.65836$
Index of Canada balsam.....	$n_B = 1.55$
Index of E ray.....	$n_E = 1.48641$

Optically the balsam is more dense than the calcite for the E ray, and less dense for the O ray. The E ray therefore will be refracted into the balsam and on through the calcite crystal, whereas the O ray for large angles of incidence will be totally reflected. The critical angle for total reflection of the O ray at the first calcite to balsam surface is about 69° and corresponds to a limiting angle SMS_0 in Fig. 24*N* of about 14° . At greater angles than this, some of the O ray will be transmitted. This means that a nicol should not be used in light which is highly convergent or divergent.

The E ray in a nicol also has an angular limit, beyond which it will be totally reflected by the balsam. This is due to the fact that the index of refraction of calcite is different for different directions through the crystal. In the next chapter it will be seen that the index $n_E = 1.486$, as it is usually given, is just for the special case of light traveling at right angles to the optic axis. Along the optic axis the E ray travels with the same speed as the O ray and it therefore has the same index of 1.658. For intermediate angles the effective index lies between these two limits 1.486 and 1.658. There will therefore be a maximum angle SMS_E beyond which the balsam will be optically less dense than the calcite, and there will be total reflection of the E vibrations. The prism is so cut that this angle likewise is in the neighborhood of 14° . The

* Very complete descriptions of polarizing prisms will be found in A. Johanssen, "Manual of Petrographic Methods," 2d ed., pp. 158-164, McGraw-Hill Book Company, Inc., New York, 1918.

direction of the incident light on a nicol therefore is limited on the one side to avoid having the O ray transmitted and on the other side to avoid having the E ray totally reflected. In practice, it is important to keep this limitation in mind.

Polarizing prisms are sometimes made with end faces cut perpendicular to the sides so that the light enters and leaves normal to the surface. The most popular one of this type, the *Glan-Thompson prism*, has an angular tolerance or aperture approaching 40° , hence much larger than that of the nicol. But this prism must be cut with the optic axis parallel to the end faces and is wasteful of calcite, large crystals of which are expensive and difficult to obtain. In another type the halves are held together so that there is a film of air between them instead of balsam. This device, called the *Foucault prism*, will transmit ultraviolet light. It has

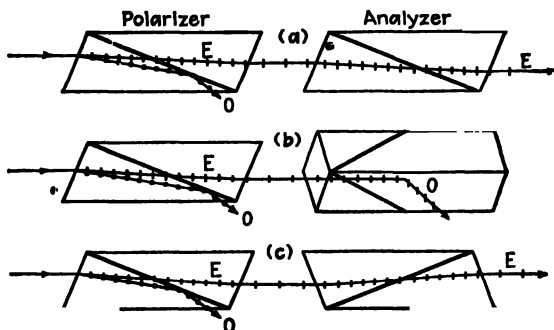


FIG. 24O. Two nicol prisms mounted as polarizer and analyzer.

an angular aperture of only about 8° , however, and some difficulty is experienced with interference occurring in the air film.

24.12. Parallel and Crossed Nicols. When two nicol prisms are lined up one behind the other, as shown in Fig. 24O, they form a good polariscope (Sec. 24.4). Positions (a) and (c) are referred to as "parallel nicols," and for them the E ray is transmitted. A loss of some 10 per cent of the incident light is caused by reflection at the prism faces and absorption in the balsam layer, so that the over-all transmission of a nicol for incident unpolarized light is about 40 per cent. Position (b) in the figure represents one of the two positions called "crossed nicols." Here the E ray from the first nicol becomes an O ray in the second, and is totally reflected to the side. For intermediate angles, the incident E vibrations from the first nicol are broken up into components as shown by the vector diagram in Fig. 24M, where θ is the angle between the principal sections of the two nicols. The E' component is transmitted

by the second nicol with the intensity $E^2 \cos^2 \theta$ and the O'' component is totally reflected.

24.13. Refraction by Calcite Prisms. Calcite prisms are sometimes cut from crystals for the purpose of illustrating double refraction and dispersion simultaneously as well as single refraction along the optic axis. Two regular prisms of calcite are shown in Fig. 24P, the first cut with the optic axis parallel to the *base* and the refracting edge *A*, and the other with the axis also parallel to the base but perpendicular to the refracting edge. In the first prism there is double refraction for all wavelengths and hence two complete spectra of plane-polarized light, one with the electric vector parallel to the plane of incidence and the other with the electric vector perpendicular to it. An interesting demonstration of this

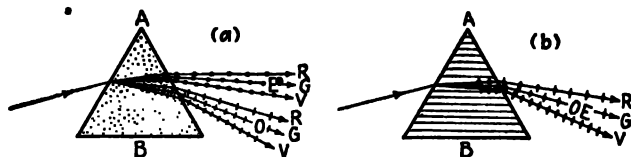


FIG. 24P. Double and single refraction of white light by prisms cut from calcite crystals.

polarization is accomplished by inserting a polarizer* into the incident or refracted beams. Upon rotation of the polarizer, first one spectrum is extinguished and then the other.

In the second prism Fig. 24P(b) only one spectrum is observed as in the case with glass prisms. Here the light travels along the optic axis, or very nearly so, so that the two spectra are superposed. In this case a polarizer, when rotated, will not affect the intensity as it does with the first prism. The more detailed treatment of double refraction in the next chapter will clarify these experimental observations.

24.14. Rochon and Wollaston Prisms. Nicol prisms cannot be used in ultraviolet light, as the Canada balsam is not transparent to these shorter wavelengths. For this purpose other types of prisms have been designed, the most satisfactory of which are the Rochon or Wollaston prisms. These optical devices, sometimes called double-image prisms, are made of quartz or calcite, cut at certain definite angles and cemented together with glycerine or castor oil.

In the Rochon prism [Fig. 24Q(a)] the light, entering normal to the surface, travels along the optic axis of the first prism and then undergoes

* Although nicol prisms are perhaps the best polarizing devices found in most laboratories, polaroid films or a pile of glass plates mounted as in Fig. 24G are quite suitable for nearly all experimental demonstrations.

double refraction at the boundary of the second prism as shown. The optic axis of the second prism is perpendicular to the plane of the page, as is indicated by the dots. In the Wollaston prism [Fig. 24Q(b)] the light enters normal to the surface and travels perpendicular to the optic axis until it strikes the second prism, where double refraction takes place. The essential difference between the two is shown in the figures by the directions of the two refracted rays. The Rochon prism transmits the *O* vibrations without deviation, the beam being achromatic. This is frequently desired in optical instruments where only one plane-polarized beam is desired. The *E* beam, which is chromatic, is readily screened off at a sufficiently large distance from the prism.

The Wollaston prism deviates both rays and consequently yields greater separation of the two chromatic beams. This device is par-

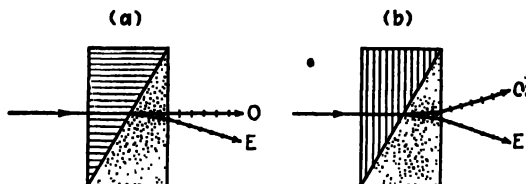


FIG. 24Q. Diagrams of (a) Rochon and (b) Wollaston prisms made from quartz.

ticularly useful where intensities in polarized light are involved, since the images of the two beams, whose vibrations are perpendicular to each other, can be compared side by side. It should be noted that in the Rochon prism the light should always enter from the left in order to travel first along the optic axis as shown in the figure. If sent through in the other direction, the different wavelengths will come out vibrating in different planes owing to a phenomenon called *rotatory dispersion* (see Sec. 27.1). This phenomenon as well as the directions taken by doubly refracted beams in quartz will be treated in detail in the following three chapters.

24.15. Polarization by Scattering. The scattering of light by small ultramicroscopic particles has been discussed in some detail in Sec. 22.9. There it was stated that for particles smaller than a wavelength of light the scattering is proportional to the fourth power of the frequency. While this law accounts for the blue color of the sky, it may also be shown to account for the polarization of the blue light as well as for the orange and red color of the sun at sunset.

The polarization of scattered light follows directly from the classical picture of light as a transverse wave with no longitudinal component. If we consider, for example, ordinary light incident on a charged par-

ticle e in Fig. 24R, the light scattered in any direction in the transverse plane must be plane-polarized with the electric vector as shown. In all other directions the scattered light will be only partially polarized (see the last two paragraphs of Sec. 22.12 and Fig. 22H). The reason why the light scattered in the transverse plane has no component vibration parallel to the direction of propagation of the incident light is that the incident light, being a transverse wave, can have no longitudinal component.

Consider now the blue light of a clear sky as seen by an observer at the time near sunset. Let Q represent the observer on the earth with

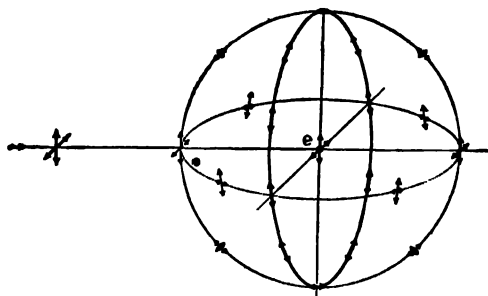


FIG. 24R. Polarization by scattering from a single particle.

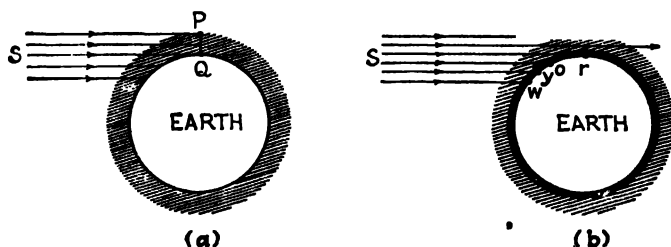


FIG. 24S. The scattering of light by the earth's atmosphere.

the sunlight incident through the atmosphere as shown in Fig. 24S(a). Looking straight overhead, the light scattered from particles in the neighborhood of P will be partially plane-polarized with a maximum perpendicular to the plane SQP . The reason pure plane-polarized light is not observed is that multiple scattering occurs in the relatively long light paths. By multiple scattering is meant that many of the incident light waves have been scattered several times before reaching the observer. Maximum polarization of this light, in agreement with theory, is observed in a direction perpendicular to the incident sunlight.

The frequent observation of a red sunset is readily attributed to the

scattering of light by fine dust and smoke particles near the earth's surface. This is illustrated by the double shaded area in Fig. 24S(b). If an observer is at the point marked (*w*) it is midafternoon, and the direct sunlight travels a relatively short dust path. Here only a little of the blue and violet have been lost by scattering and the sun appears white. To an observer at (*y*) the dust path is increased with the result that most of the blue and violet have been scattered and the sun, owing to the remaining colors red, orange, yellow, and green, appears yellow at (*o*) the blue, violet, and a good share of the green are scattered and the sun appears orange since now only the red, orange, and yellow reach the eye. At (*r*) it is sunset, and the dust path has increased to many times that for (*w*). Along this path all but a little of the very longest visible waves are scattered and the sun appears red. Looking overhead, the sky still appears blue and the light from it is partially plane-polarized.

A very interesting experimental demonstration of the blue sky and red sunset may be performed as shown in Fig. 24T'. Light from a bright

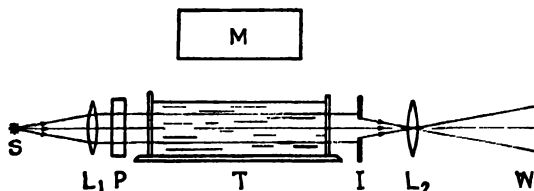


FIG. 24T'. Experimental arrangement for demonstrating polarization by scattering, and for showing the origin of the blue sky and red sunset.

source *S* (sunlight or a carbon arc) is made to pass through a lens *L*₁, a tank of clear water *T*, a screen with a circular aperture *I*, a lens *L*₂, and finally to a screen *W*. Lens *L*₁ produces a parallel beam of light in the tank, and *L*₂ focuses an image of the circular opening *I* on the screen. About 20 g of photographic fixing powder (hyposulfite of soda) are dissolved in each gallon of water in the tank *T*. This tank should be from 12 to 24 in. long. If about 1 to 2 cc of concentrated sulfuric acid is now added to the tank, fine microscopic sulfur particles will begin to precipitate slowly. The correct amount of salt and acid to produce the best effect must be determined by trial. It should take 2 or 3 min for the first visible precipitate to form.

As the particles begin to form, scattered blue light will mark the path of the beam of light through the tank. Viewed through a nicol prism or other analyzer, from a direction at right angles to the light beam, the light is first found to be plane-polarized and later, after more particles

have formed, to be partially polarized, as predicted. The bright circular image on the screen, representing the sun, will be observed to change slowly from white to yellow then to orange and finally red. In the latter stages of the experiment multiple scattering causes the whole front end of the tank to be blue. The other end is yellow and orange since the blue and violet has been scattered out of the beam. When demonstrated to an audience where individual analyzers are not to be had, a large nicol or polarizing plate can be inserted at the position P . Upon rotating this polarizer the scattering is made to appear or disappear with each 90° rotation. A large mirror M placed directly above the tank will in this case show the beam alternately in the mirror and in the tank.

In performing an experiment of this kind Tyndall was the first to observe another type of scattering when the particles become large enough to scatter white light. If the white scattered light is viewed through a nicol held in the position for the ordinary extinction of blue scattered light, the blue color appears again with increased brilliancy. This Tyndall called the residual blue. A theoretical consideration of the phenomenon by Rayleigh shows that this scattering is proportional to ν^4 . With smaller particles the residual blue vanishes and a zero minimum is observed.

Problems

1. A beam of light traveling in water strikes the surface of a glass plate, and when the angle of incidence is adjusted to be 50.82° the reflected beam is found to be plane-polarized. What is the refractive index of the glass?
2. Calculate the polarizing angle for a dense flint glass of index $n = 1.768$.
3. Calculate the limits of the polarizing angle for white light, $\lambda = 4000$ to 7000 Å for telescope crown glass as given by the dispersion curve in Fig. 23A.
4. Prove that when light is incident on a plane-parallel glass plate at the polarizing angle for the upper surface, the refracted beam also meets the lower surface at the polarizing angle for that surface.
5. A beam of plane-polarized light is incident normally on a calcite crystal with its vibrations making an angle of 17° with the principal section. Calculate the relative amplitudes and intensities of the two refracted beams.
6. Find the relative intensities of the four images in Fig. 24L(b) if the angle between the principal sections of the two crystals is 9.5° .
7. A beam of plane-polarized light is incident normally on a calcite crystal with its vibrations making an angle of 26° with the principal section. The two refracted beams now pass through a nicol prism placed behind the calcite and oriented with its principal section at 70° with the original light vibrations and at 44° with the principal section. Calculate the relative intensities of the two beams.
8. From the index of refraction of calcite and Canada balsam, compute the maximum angle at which light may be incident on a nicol prism and still have the ordinary ray totally reflected by the Canada balsam ($n_o = 1.6584$, $n_{CB} = 1.55$). What does

this give for the maximum allowable angle with respect to the axis of the nicol (angle S_0MS in Fig. 24N)?

9. If two nicol prisms are mounted as a polarizer and an analyzer with their principal sections making an angle of 16° with each other, what will be the relative intensity of the transmitted light when the angle is changed to 46° ?

10. Two light sources are observed, one after the other, with two nicol prisms mounted one behind the other as polarizer and analyzer. What are the relative intensities of the two sources if the intensities of the emergent beams in both instances are equal when the angles between the principal planes of the nicols are 45° and 70° , respectively?

11. Ordinary light is incident on a pile of eight glass plates at the polarizing angle. If $n = 1.602$, find the proportion of polarization of the light transmitted.

12. Four polaroid sheets are placed on top of each other, each with its axis rotated by a certain angle α with respect to the preceding one. Find the transmitted intensity, relative to that of the incident unpolarized light, (a) when $\alpha = 4^\circ$, and (b) when $\alpha = 32^\circ$. Assume that because of absorption any one sheet transmits only 40 per cent of incident unpolarized light (80 per cent of incident light polarized parallel to its plane of its transmission). What is the orientation of the emergent vibrations in cases (a) and (b)?

13. Quartz has $n_o = 1.54425$ and $n_E = 1.55336$ for sodium light. Calculate the angle between the emergent O and E rays for a Rochon prism (Fig. 24Q) made of quartz with prism angles of 20° . (NOTE: As will be shown in the next chapter, the refractive index for the E ray has the value n_o when the ray is parallel to the optic axis and the value n_E when it is perpendicular to this axis.)

14. Calculate the percentage polarization of the light scattered by very small particles at an angle of 70° with the direction of the incident unpolarized beam. Any multiple scattering is, of course, to be neglected.

15. Given that the three angles at the blunt corner of a calcite crystal are $101^\circ 54'$, find the angles (a) between each pair of faces forming this corner, (b) between the optic axis and an edge where the faces form an obtuse angle, and (c) between the optic axis and the face of the crystal, in the principal section.

16. Find the number of glass plates of index $n = 1.55$ which must be used in a pile of plates to obtain 90 per cent polarization of the transmitted light.

